การเลือกคุณลักษณะสำคัญแบบพลวัตสำหรับความเหมาะสมที่สุดของการจำแนกต้นไม้ตัดสินใจบนพื้นฐานเงื่อนไขหลายเป้าหมาย

นางสาวนิชาภา จำปาศรี

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรมหาบัณฑิต
แขนงวิชาเทคโนโลยีสารสนเทศและการสื่อสาร มหาวิทยาลัยสุโขทัยธรรมาธิราช
พ.ศ. 2563

Dynamic Feature Selection for Optimization of Decision Tree Classification Based on Multi-target Conditions

Miss Nichapa Jampasri

A Thesis Submitted in Partial Fulfillment of the Requirements for
the Degree of Master of Science in Information and Communication Technology
School of Science and Technology
Sukhothai Thammathirat Open University
2020

หัวข้อวิทยานิพนธ์ การเลือกคุณลักษณะสำคัญแบบพลวัตสำหรับความเหมาะสมที่สุดของการจำแนกต้นไม้ตัดสินใจบนพื้นฐานเงื่อนไขหลายเป้าหมาย

ชื่อและนามสกุล นางสาวนิชาภา จำปาศรี

แขนงวิชา เทคโนโลยีสารสนเทศและการสื่อสาร

สาขาวิชา วิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยสุโขทัยธรรมาธิราช

อาจารย์ที่ปรึกษา 1. รองศาสตราจารย์ ดร.วฤษาย์ ร่มสายหยุด

2. อาจารย์ ดร.เอกสิทธิ์ พัชรวงศ์ศักดา

วิทยานิพนธ์นี้ ได้รับความเห็นชอบให้เป็นส่วนหนึ่งของการศึกษา
ตามหลักสูตรระดับปริญญาโท เมื่อวันที่ 23 มิถุนายน 2564

คณะกรรมการสอบวิทยานิพนธ์

..... ประธานกรรมการ
(ผู้ช่วยศาสตราจารย์ ดร.วราภรณ์ วิยานนท์)

..... กรรมการ
(รองศาสตราจารย์ ดร.วฤษาย์ ร่มสายหยุด)

..... กรรมการ
(อาจารย์ ดร.เอกสิทธิ์ พัชรวงศ์ศักดา)

..... ประธานกรรมการบัณฑิตศึกษา
(รองศาสตราจารย์ ดร.เทพศักดิ์ บุณยรัตพันธุ์)

ชื่อวิทยานิพนธ์ การเลือกคุณลักษณะสำคัญแบบพลวัตสำหรับความเหมาะสมที่สุดของการจำแนกต้นไม้ตัดสินใจบนพื้นฐานเงื่อนไขหลายเป้าหมาย

ผู้วิจัย นางสาวนิชาภา จำปาศรี **รหัสนักศึกษา** 2619600014

ปริญญา วิทยาศาสตรมหาบัณฑิต (เทคโนโลยีสารสนเทศและการสื่อสาร)

อาจารย์ที่ปรึกษา (1) รองศาสตราจารย์ ดร.วฤษาย์ ร่มสายหยุด

(2) อาจารย์ ดร.เอกสิทธิ์ พัชรวงศ์ศักดา

ปีการศึกษา 2563

บทคัดย่อ

การวิจัยครั้งนี้มีวัตถุประสงค์ 1) เพื่อบูรณาการอัลกอริทึมการเลือกคุณลักษณะสำคัญแบบพลวัตกับการจำแนกต้นไม้ตัดสินใจบนพื้นฐานเงื่อนไขหลายเป้าหมาย และ 2) เพื่อประเมินประสิทธิภาพการเลือกคุณลักษณะสำคัญแบบพลวัตสำหรับความเหมาะสมที่สุดของการจำแนกต้นไม้ตัดสินใจบนพื้นฐานเงื่อนไขหลายเป้าหมาย

การดำเนินงานประกอบด้วย 6 ขั้นตอนหลัก ได้แก่ 1) การเก็บรวบรวมข้อมูลของนักศึกษาจากงานทุนการศึกษาของวิทยาลัยเทคโนโลยีทางการแพทย์และสาธารณสุข กาญจนาภิเษก ข้อมูลจำนวน 500 คน และคุณลักษณะสำคัญจำนวน 29 คุณลักษณะ 2) การเตรียมข้อมูลโดยกำหนดเงื่อนไขหลายเป้าหมายจำนวน 3 แบบ สำหรับทุนที่มี 3 ประเภท โดยประยุกต์วิธีสังเคราะห์ข้อมูลเพิ่มสำหรับแก้ไขปัญหาชุดข้อมูลของนักศึกษาที่มีความไม่สมดุล และพัฒนาวิธีการเลือกคุณลักษณะสำคัญแบบพลวัตบนพื้นฐานเงื่อนไขหลายเป้าหมาย 3) การสร้างแบบจำลองการจำแนกด้วยอัลกอริทึมต้นไม้ตัดสินใจ สำหรับสอนและทดสอบแบบจำลอง 4) การประเมินประสิทธิภาพแบบจำลอง 5) การปรับค่าพารามิเตอร์เพื่อหาค่าความเหมาะสมที่สุด และ 6) การใช้โมเดลพยากรณ์

ผลการวิจัยนี้ได้ผลลัพธ์ค่าความถูกต้องร้อยละ 85.37 ค่าความแม่นยำร้อยละ 85.12 ค่าเรียกคืนร้อยละ 86.52 และการวัดประสิทธิภาพโดยรวมร้อยละ 85.79

คำสำคัญ การเลือกคุณลักษณะสำคัญแบบพลวัต, การจำแนกประเภท, ต้นไม้ตัดสินใจ, เงื่อนไขหลายเป้าหมาย

Thesis title: Dynamic Feature Selection for Optimization of Decision Tree Classification Based on Multi-target Conditions

Researcher: Miss Nichapa Jampasri; **ID:** 2619600014;

Degree: Master of Science (Information and Communication Technology);

Thesis advisors: (1) Dr.Walisa Romsaiyud, Associate Professor;

(2) Dr.Eakasit Pacharawongsakda;

Academic year: 2020

Abstract

The purposes of this research were 1) to integrate the dynamic feature selection algorithm with the decision tree classification based on multi-target conditions, and 2) to evaluate the performance of the dynamic feature selection for optimization of decision tree classification based on multi-target conditions.

The operation consisted of six main steps as 1) data collection of student scholarships from Kanchanabhishek Institute of Medical and Public Health Technology on 500 students and 29 features, 2) data preparation by assigning the multi-target conditions for 3 scholarship types by applying the synthetic minority over-sampling technique for solving the problem with imbalanced dataset and developing the dynamic feature selection algorithm based on multi-target conditions, 3) building the classification model with decision tree algorithm for training and testing a model, 4) evaluation the model, 5) parameter tuning for finding the optimal value and 6) model prediction.

The experimental results were 85.37% for accuracy, 85.12 % for precision, 86.52 % for recall, and 85.79% for F-measure.

Keywords: Dynamic feature selection, Classification, Decision Tree, Multi-target Conditions

กิตติกรรมประกาศ

วิทยานิพนธ์ฉบับนี้ สำเร็จลุล่วงไปได้โดยได้รับความอนุเคราะห์ข้อมูลจากวิทยาลัยเทคโนโลยีทางการแพทย์และสาธารณสุข กาญจนาภิเษก ผู้วิจัยขอขอบพระคุณเป็นอย่างสูง และด้วยความกรุณาและความช่วยเหลือเป็นอย่างดียิ่งจากรองศาสตราจารย์ ดร.วฤษาย์ ร่มสายหยุด อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก อาจารย์ ดร.เอกสิทธิ์ พัชรวงศ์ศักดา อาจารย์ที่ปรึกษาร่วม และและผู้ช่วยศาสตราจารย์ ภิรมย์ คงเลิศ ที่กรุณาเสียสละเวลาให้คำแนะนำคำปรึกษา ตรวจแก้ไขข้อบกพร่องด้วยความเอาใจใส่ ผู้วิจัยรู้สึกซาบซึ้งในความกรุณาของอาจารย์และขอกราบขอบพระคุณเป็นอย่างสูงไว้ ณ ที่นี้

ขอขอบคุณบุคลากร อาจารย์ และเจ้าหน้าที่ จากวิทยาลัยเทคโนโลยีทางการแพทย์และสาธารณสุข กาญจนาภิเษก ที่ให้ความช่วยเหลือ คำแนะนำ สนับสนุน และคอยให้คำปรึกษาตลอดระยะเวลาการศึกษาในครั้งนี้

สุดท้ายนี้ ขอกราบขอบพระคุณบิดา มารดา และครอบครัวที่เป็นกำลังใจและสนับสนุนให้การช่วยเหลือทำให้งานวิจัยฉบับนี้สำเร็จลุล่วงไปได้ด้วยดี คุณประโยชน์อันเกิดจากวิทยานิพนธ์ฉบับนี้ขอน้อมบูชาแด่พระคุณบิดา มารดา และครูอาจารย์ที่คอยอบรมสั่งสอน ให้คำแนะนำ สนับสนุนและให้กำลังใจอย่างดียิ่งแก่ผู้วิจัย หวังเป็นอย่างยิ่งว่าการศึกษาครั้งนี้คงจะเป็นประโยชน์แก่ผู้นำไปศึกษาค้นคว้าเพื่อเป็นแนวทางในการดำเนินการของหน่วยงานที่เกี่ยวข้องให้มีประสิทธิภาพมากขึ้น หากมีข้อผิดพลาดประการใด ผู้วิจัยขออภัยไว้ ณ ที่นี้ด้วย

นิชาภา จำปาศรี

มิถุนายน 2564

สารบัญ

หน้า

บทคัดย่อภาษาไทย ..... ง
บทคัดย่อภาษาอังกฤษ ..... จ
กิตติกรรมประกาศ ..... ฉ
สารบัญ ..... ช
สารบัญตาราง ..... ญ
สารบัญภาพ ..... ฎ
บทที่ 1 บทนำ 1
ความเป็นมาและความสำคัญของปัญหา 1
วัตถุประสงค์การวิจัย 3
ขอบเขตของการวิจัย 3
นิยามศัพท์เฉพาะ 3
ประโยชน์ที่คาดว่าจะได้รับ 5
บทที่ 2 วรรณกรรมที่เกี่ยวข้อง 6
แนวคิดและทฤษฎี 7
การเรียนรู้ของเครื่อง (Machine Learning : ML) 7
การคัดเลือกคุณลักษณะสำคัญ (feature selection) 12
การคัดเลือกคุณลักษณะสำคัญแบบพลวัต (dynamic feature selection) 16
การจัดการความไม่สมดุลของข้อมูล (imbalanced data) 18
การสุ่มตัวอย่างกลุ่มน้อยสังเคราะห์ (Synthetic Minority Oversampling Technique : SMOTE) 18
เทคนิคการจำแนกประเภทข้อมูล (classification) 19
เทคนิคต้นไม้ตัดสินใจ (decision tree) 20
การวัดประสิทธิภาพแบบจำลอง (evaluation) 21
งานวิจัยที่เกี่ยวข้อง 24
งานวิจัยในประเทศ 24
งานวิจัยต่างประเทศ 27

สารบัญ (ต่อ)

หน้า

บทที่ 3 วิธีดำเนินการวิจัย 37
ประชากรและกลุ่มตัวอย่าง 37
เครื่องมือที่ใช้ในการวิจัย 37
ขั้นตอนการดำเนินการวิจัย 38
การเก็บรวบรวมข้อมูล (data collection) 39
การเตรียมข้อมูล (data preparation) 39
การสร้างแบบจำลอง (modeling) 40
การวัดประสิทธิภาพโมเดล (evaluation the model) 40
การปรับค่าพารามิเตอร์ (parameter tuning) 40
การใช้งานการทำนายแบบจำลอง (model prediction) 40
บทที่ 4 ผลการดำเนินการวิจัย 41
การเก็บรวบรวมข้อมูล (data collection) 41
การเตรียมข้อมูล (data preparation) 43
การจัดการความไม่สมดุลของข้อมูล (imbalanced data classification) 48
การเลือกคุณลักษณะสำคัญ (feature selection) 52
การเลือกคุณลักษณะสำคัญแบบพลวัต (dynamic feature selection) 61
การสร้างแบบจำลอง (modeling) 69
การวัดประสิทธิภาพโมเดล (evaluation the model) 82
การปรับค่าพารามิเตอร์ (parameter tuning) 83
การใช้งานการทำนายแบบจำลอง (model prediction) 83
บทที่ 5 สรุปผลการวิจัย อภิปรายผล และข้อเสนอแนะ 84
สรุปการวิจัย 84
อภิปรายผล 85
ข้อเสนอแนะ 86

สารบัญ (ต่อ)

หน้า

บรรณานุกรม....87
ภาคผนวก....89
ภาคผนวก ก หนังสือขอความอนุเคราะห์ให้นักศึกษาเก็บข้อมูลเพื่อการศึกษาวิทยานิพนธ์....90
ภาคผนวก ข เอกสารรับรองโครงการวิจัย....92
ภาคผนวก ค ชุดข้อมูล (dataset)....94
ประวัติผู้วิจัย....103

สารบัญตาราง

หน้า

ตารางที่ 2.1 ตารางจำแนกประเภทของอัลกอริทึมแบบ supervised และ unsupervised learning..... 9
ตารางที่ 2.2 ตาราง confusion matrix..... 22
ตารางที่ 2.3 ตารางสรุปเปรียบเทียบงานวิจัยที่เกี่ยวข้อง..... 30
ตารางที่ 2.4 ตารางเปรียบเทียบรายละเอียดงานวิจัยที่เกี่ยวข้อง..... 32
ตารางที่ 4.1 แสดงรายละเอียดคุณลักษณะของข้อมูล..... 42
ตารางที่ 4.2 แสดงการแปลงรูปข้อมูลนักศึกษาให้อยู่ในรูปแบบที่พร้อมสำหรับนำไปใช้วิเคราะห์ข้อมูล..... 44
ตารางที่ 4.3 แสดงข้อมูลที่มีความไม่สมดุล (imbalanced data)..... 48
ตารางที่ 4.4 แสดงข้อมูลที่ผ่านการปรับสมดุลข้อมูลด้วยวิธี SMOTE..... 51
ตารางที่ 4.5 แสดงค่าความถูกต้องของเทคนิค Correlation base Feature Selection..... 53
ตารางที่ 4.6 แสดงค่าความถูกต้องของเทคนิค Information Gain..... 54
ตารางที่ 4.7 แสดงค่าความถูกต้องของเทคนิค Gain Ratio..... 55
ตารางที่ 4.8 แสดงค่าความถูกต้องของเทคนิค Chi Square..... 56
ตารางที่ 4.9 แสดงค่าความถูกต้องของเทคนิค Forward Selection..... 57
ตารางที่ 4.10 แสดงค่าความถูกต้องของเทคนิค Backward Elimination..... 58
ตารางที่ 4.11 แสดงค่าความถูกต้องของเทคนิค Evolutionary Selection..... 59
ตารางที่ 4.12 แสดงค่าความถูกต้องของเทคนิคการเลือกคุณลักษณะสำคัญ..... 60
ตารางที่ 4.13 แสดงค่าน้ำหนักของคุณลักษณะของเป้าหมาย A..... 65
ตารางที่ 4.14 แสดงค่าน้ำหนักของคุณลักษณะของเป้าหมาย B..... 66
ตารางที่ 4.15 แสดงค่าน้ำหนักของคุณลักษณะของเป้าหมาย C..... 67
ตารางที่ 4.16 แสดงค่าน้ำหนักของแต่ละคุณลักษณะที่ได้จากการเลือกคุณลักษณะสำคัญแบบพลวัต..... 68
ตารางที่ 4.17 แสดงผลการเลือกคุณลักษณะสำคัญแบบพลวัต..... 69
ตารางที่ 4.18 แสดงผลการวัดประสิทธิภาพของแบบจำลองด้วยชุดข้อมูล 100 แถว..... 72
ตารางที่ 4.19 แสดงผลการวัดประสิทธิภาพของแบบจำลองด้วยชุดข้อมูล 300 แถว..... 73
ตารางที่ 4.20 แสดงผลการวัดประสิทธิภาพของแบบจำลองด้วยชุดข้อมูล 500 แถว..... 74
ตารางที่ 4.21 แสดงผลการเปรียบเทียบค่าความถูกต้องของแบบจำลอง..... 75
ตารางที่ 4.22 แสดงผลการวัดประสิทธิภาพของแบบจำลอง..... 82

สารบัญภาพ

หน้า

ภาพที่ 2.1 แสดงตัวอย่างการกำหนดคลาสเป้าหมายแบบไบนารีในรูปแบบ single label 16
ภาพที่ 2.2 แสดงตัวอย่างการกำหนดคลาสเป้าหมายแบบหลายค่าในรูปแบบของ single label 17
ภาพที่ 2.3 แสดงตัวอย่างการกำหนดคลาสเป้าหมายแบบไบนารี (binary) ในรูปแบบ multi-label 17
ภาพที่ 2.4 แสดงความไม่สมดุลของข้อมูล (imbalanced data)..... 18
ภาพที่ 2.5 แสดงการปรับสมดุลของข้อมูลด้วยวิธี SMOTE..... 19
ภาพที่ 2.6 แสดงตัวอย่างต้นไม้ตัดสินใจที่ใช้ในการจำแนก 20
ภาพที่ 2.7 แสดงการทดสอบแบบไขว้ทบ (k – fold cross validation)..... 22
ภาพที่ 3.1 แสดงขั้นตอนการเลือกคุณลักษณะสำคัญแบบพลวัตสำหรับความเหมาะสมที่สุดของการจำแนกต้นไม้ตัดสินใจบนพื้นฐานเงื่อนไขหลายเป้าหมาย..... 38
ภาพที่ 4.1 แสดงตัวอย่างข้อมูลที่ไม่สอดคล้อง (inconsistent data) และข้อมูลที่มีค่าผิดพลาด (error)..... 43
ภาพที่ 4.2 แสดงตัวอย่างข้อมูลที่มีค่าผิดพลาด (error) 44
ภาพที่ 4.3 แสดงภาพรวมกระบวนการนำเข้าข้อมูลและการปรับสมดุลของข้อมูลด้วยวิธี SMOTE .. 49
ภาพที่ 4.4 แสดงการนำเข้าข้อมูล (training data) 49
ภาพที่ 4.5 แสดงการเลือกคุณลักษณะของข้อมูล 50
ภาพที่ 4.6 แสดงการกำหนดเป้าหมาย 50
ภาพที่ 4.7 แสดงการกำหนดจำนวนข้อมูล sample..... 50
ภาพที่ 4.8 แสดงการปรับสมดุลข้อมูลด้วยวิธี SMOTE 51
ภาพที่ 4.9 แสดงภาพรวมกระบวนการเลือกเทคนิคที่เหมาะสมสำหรับการเลือกคุณลักษณะสำคัญแบบพลวัต..... 52
ภาพที่ 4.10 แสดงการออกแบบการคัดเลือกคุณลักษณะสำคัญของเทคนิค Correlation base Feature Selection 53
ภาพที่ 4.11 แสดงการออกแบบการคัดเลือกคุณลักษณะสำคัญของเทคนิค Information Gain..... 54
ภาพที่ 4.12 แสดงการออกแบบการคัดเลือกคุณลักษณะสำคัญของเทคนิค Gain Ratio..... 55
ภาพที่ 4.13 แสดงการออกแบบการคัดเลือกคุณลักษณะสำคัญของเทคนิค Chi Square 56
ภาพที่ 4.14 แสดงการออกแบบการคัดเลือกคุณลักษณะสำคัญของเทคนิค Forward Selection... 57
ภาพที่ 4.15 แสดงการออกแบบการคัดเลือกคุณลักษณะสำคัญของเทคนิค Backward Elimination 58

สารบัญภาพ (ต่อ)

หน้า

ภาพที่ 4.16 แสดงการออกแบบการคัดเลือกคุณลักษณะสำคัญของเทคนิค Evolutionary Selection 59
ภาพที่ 4.17 แสดงภาพรวมกระบวนการเลือกคุณลักษณะสำคัญแบบพลวัตกับการจำแนกต้นไม้ตัดสินใจ 61
ภาพที่ 4.18 แสดงการนำเข้าข้อมูล (training data) 62
ภาพที่ 4.19 แสดงการเลือกคุณลักษณะของข้อมูล 62
ภาพที่ 4.20 แสดงการกำหนดเป้าหมาย 63
ภาพที่ 4.21 แสดงการกำหนดจำนวนข้อมูล sample 63
ภาพที่ 4.22 แสดงการปรับสมดุลข้อมูลด้วยวิธี SMOTE 63
ภาพที่ 4.23 แสดงการกำหนดค่าพารามิเตอร์ select by weights 63
ภาพที่ 4.24 แสดงการกำหนด optimize parameters (grid) 64
ภาพที่ 4.25 แสดงการกำหนด cross validation เท่ากับ 10 64
ภาพที่ 4.26 แสดงภาพรวมกระบวนการสร้างแบบจำลองการจำแนกประเภทข้อมูลด้วยเทคนิคต้นไม้ตัดสินใจ (decision tree) 69
ภาพที่ 4.27 แสดงการกำหนดเป้าหมายโดยใช้ set role2 70
ภาพที่ 4.28 แสดงการกำหนด cross validation เท่ากับ 10 70
ภาพที่ 4.29 แสดงการสร้างแบบจำลองการจำแนกต้นไม้ตัดสินใจ 71
ภาพที่ 4.30 แสดงผลการเปรียบเทียบค่าความถูกต้องโดยใช้ชุดข้อมูล 100 72
ภาพที่ 4.31 แสดงผลการเปรียบเทียบค่าความถูกต้องโดยใช้ชุดข้อมูล 300 73
ภาพที่ 4.32 แสดงผลการเปรียบเทียบค่าความถูกต้องโดยใช้ชุดข้อมูล 500 74
ภาพที่ 4.33 แสดงผลการเปรียบเทียบค่าความถูกต้องโดยใช้ชุดข้อมูล 100, 300 และ 500 75
ภาพที่ 4.34 แสดงภาพรวมการสร้างแบบจำลองการจำแนกประเภทต้นไม้ตัดสินใจของเป้าหมาย A 76
ภาพที่ 4.35 แสดงผลการสร้างแบบจำลองการจำแนกประเภทต้นไม้ตัดสินใจของเป้าหมาย A 77
ภาพที่ 4.36 แสดงภาพรวมการสร้างแบบจำลองการจำแนกประเภทต้นไม้ตัดสินใจของเป้าหมาย B 78
ภาพที่ 4.37 แสดงผลการสร้างแบบจำลองการจำแนกประเภทต้นไม้ตัดสินใจของเป้าหมาย B 79
ภาพที่ 4.38 แสดงภาพรวมการสร้างแบบจำลองการจำแนกประเภทต้นไม้ตัดสินใจของเป้าหมาย C 80
ภาพที่ 4.39 แสดงผลการสร้างแบบจำลองการจำแนกประเภทต้นไม้ตัดสินใจของเป้าหมาย C 81
ภาพที่ 4.40 แสดงผลการเปรียบเทียบการวัดประสิทธิภาพของแบบจำลอง 82
ภาพที่ 4.41 แสดงการปรับค่าพารามิเตอร์ของแบบจำลอง 83

บทที่ 1
บทนำ

1. ความเป็นมาและความสำคัญของปัญหา

ปัจจุบันมีการประยุกต์ใช้การเรียนรู้ของเครื่อง (Machine Learning: ML) (Ethem, 2020) ในงานหลายประเภททั้งในด้านอุตสาหกรรม ธนาคาร วิทยาศาสตร์และการแพทย์ เพื่อส่งเสริมเศรษฐกิจ และคุณภาพชีวิตที่ดีขึ้น สำหรับวิเคราะห์ข้อมูล พยากรณ์ข้อมูล เพื่อการส่งเสริมการตัดสินใจที่ถูกต้องและแม่นยำสูง หลักการเรียนรู้ของเครื่อง (วฤษาย์ ร่มสายหยุด, 2564) เป็นการนำข้อมูลที่มีอยู่เดิมมาเข้าสู่วิธีการหรืออัลกอริทึม (algorithms) ต่างๆ เพื่อหาความสัมพันธ์ และรูปแบบของข้อมูล เพื่อให้คอมพิวเตอร์สามารถวิเคราะห์ และตัดสินใจในการดำเนินการต่างๆ ได้เอง ส่งผลทำให้องค์กรมีกระบวนการวิเคราะห์เทคโนโลยีสารสนเทศให้ถูกต้อง และมีประสิทธิภาพ วิธีการวิเคราะห์เชิงทำนาย (predictive analytics) เป็นการนำข้อมูลที่มีอยู่เดิมในองค์กรหรือหน่วยงานต่างๆ มาทำการวิเคราะห์ด้วยวิธีการเรียนรู้ของเครื่องเพื่อทำการวิเคราะห์โอกาสหรือแนวโน้มที่จะเกิดขึ้นในอนาคตได้ จึงทำให้วิธีการเรียนรู้ของเครื่องแบบการเรียนรู้แบบมีผู้สอน (supervised learning) ถูกนำไปประยุกต์ในการทำงานอย่างกว้างขวาง และประสบความสำเร็จอย่างมาก วิทยาลัยเทคโนโลยีทางการแพทย์และสาธารณสุข กาญจนาภิเษก เป็นวิทยาลัยในสังกัดสถาบันพระบรมราชชนก สำนักงานปลัดกระทรวงสาธารณสุข ซึ่งเป็นสถาบันการศึกษาที่ผลิตบุคลากรด้านสาธารณสุข และจัดการเรียนการสอนในระดับอุดมศึกษา ได้ให้ความสำคัญกับการจัดสรรทุนการศึกษาให้กับนักศึกษาอย่างเหมาะสมที่สุด โดยมีงานทุนการศึกษาที่ทำหน้าที่หาแหล่งทุนให้กับนักศึกษาอย่างต่อเนื่อง ซึ่งการจัดสรรทุนจะพิจารณาตามหลักเกณฑ์และใช้ข้อมูลประวัติของนักศึกษามาประกอบการพิจารณา ปัจจุบันวิทยาลัยเทคโนโลยีทางการแพทย์และสาธารณสุข กาญจนาภิเษก มีทุนทั้งหมด 3 ทุน ได้แก่ ทุนได้เปล่า ทุนกู้ยืมเพื่อการศึกษา และทุนขาดแคลนทุนทรัพย์ จากเงื่อนไขการให้ทุนแก่นักศึกษานั้น นักศึกษา 1 คน สามารถยื่นขอทุนการศึกษาได้หลายทุน

ด้วยหลักการพื้นฐานของการเรียนรู้ของเครื่อง (Machine Learning: ML) และวิธีการคัดเลือกคุณลักษณะสำคัญ (feature selection) เป็นวิธีการหนึ่งที่เหมาะสมสามารถนำมาประยุกต์ใช้วิเคราะห์เชิงทำนายบนพื้นฐานเงื่อนไขหลายเป้าหมาย (multi-target) (Tanaka, E.A., et al., 2015) ได้ศึกษาปัญหาการจำแนกหลายเป้าหมาย เพื่อการทำนายผลลัพธ์ที่มากกว่าหนึ่งเป้าหมายโดยใช้เทคนิคต้นไม้ตัดสินใจ ซึ่งตัวแปรหนึ่งอาจมีความสัมพันธ์กับเป้าหมายมากกว่าหนึ่งเป้าหมายทำให้การจำแนกประเภทข้อมูลยากขึ้น มีการพิจารณาความสัมพันธ์ระหว่างเป้าหมาย พบว่า หากเป้าหมายทั้งหมดมีความสัมพันธ์กันจะได้ตัวแบบจำแนกต้นไม้ตัดสินใจเพียงตัวเดียวในการจำแนกประเภท และ

หากเป้าหมายทั้งหมดไม่มีความสัมพันธ์กันจะได้ตัวแบบจำแนกต้นไม้ตัดสินใจแยกตามเป้าหมาย ดังนั้นจึงใช้หลักการพิจารณาความเกี่ยวข้องกันของเป้าหมายการได้รับทุนประเภทต่างๆ เพื่อพัฒนาวิธีการที่เหมาะสมในสร้างโมเดลการจำแนกต้นไม้ตัดสินใจบนพื้นฐานเงื่อนไขหลายเป้าหมาย และเนื่องจากข้อมูลของนักศึกษาวิทยาลัยฯ มีความไม่สมดุลของข้อมูล (imbalanced data) เช่น ข้อมูลจำแนกประเภทกลุ่มไม่ได้รับทุนมีจำนวนมากกว่าการได้รับทุน ซึ่งทำให้เมื่อนำไปสร้างโมเดลจะทำให้ผลลัพธ์เอนเอียงไปทางกลุ่มมาก จากงานวิจัยของ (ภรันยา ปาลวิสุทธิ์, 2559) ได้ใช้เทคนิคการสุ่มเพิ่มตัวอย่างกลุ่มน้อย (SMOTE) มาแก้ไขปัญหาความไม่สมดุลของข้อมูลการเป็นโรคติดอินเทอร์เน็ต ซึ่งเทคนิคการสุ่มตัวอย่างกลุ่มน้อยสามารถเพิ่มประสิทธิภาพของแบบจำลองได้ดีขึ้น จากคุณลักษณะสำคัญของข้อมูลนักศึกษา จำนวน 29 คุณลักษณะ ทำให้ต้องมีการเลือกคุณลักษณะสำคัญ เพื่อให้ได้คุณลักษณะสำคัญที่สอดคล้องกับผลลัพธ์การทำนาย ตัวอย่างงานวิจัย (รัชพล กลัดชื่น และจรัญ แสนราช, 2561) ได้เปรียบเทียบประสิทธิภาพตัวแบบจำลองระหว่างการใช้คุณลักษณะทั้งหมดกับการคัดเลือกคุณลักษณะแบบ Forward Selection พบว่า การเลือกคุณลักษณะแบบ Forward Selection สามารถลดจำนวนคุณลักษณะลงเมื่อนำไปสร้างแบบจำลองด้วยเทคนิคต้นไม้ตัดสินใจ (decision tree แบบ j48) ให้ค่าความถูกต้องดีที่สุด (ตามรูปแบบข้อมูลของวิทยาลัยฯ) เห็นได้ว่ากระบวนการคัดเลือกคุณลักษณะ (feature selection) สามารถช่วยเพิ่มความถูกต้องในการพยากรณ์ (improving prediction accuracy) ดังงานวิจัยของ (Koller and Sahami, 1996) และ (อัจจิมา มณฑาพันธุ์, 2562) จึงทำให้มีการประยุกต์ใช้วิธีการคัดเลือกคุณลักษณะสำคัญมาใช้ในการปรับปรุงประสิทธิภาพแบบจำลองกันอย่างแพร่หลาย

ดังนั้นงานวิจัยนี้จึงนำเสนอการเลือกคุณลักษณะสำคัญแบบพลวัตสำหรับความเหมาะสมที่สุดของการจำแนกต้นไม้ตัดสินใจบนพื้นฐานเงื่อนไขหลายเป้าหมาย เพื่อเลือกคุณลักษณะสำคัญที่แปรผันไปตามเงื่อนไขหลายเป้าหมาย จากนั้นใช้เทคนิควิธีการสุ่มตัวอย่างกลุ่มน้อยสังเคราะห์ (Synthetic Minority Over-Sampling Technique: SMOTE) มาช่วยปรับสมดุลข้อมูล และเลือกใช้เทคนิคการคัดเลือกคุณลักษณะสำคัญทั้งหมด 2 ประเภท คือ 1) วิธีแบบกรอง (Filter approach) ได้แก่ เทคนิค Correlation base Feature Selection เทคนิค Information Gain เทคนิค Gain Ratio และเทคนิค Chi Square และ 2) วิธีแบบควบรวม (Wrapper approach) ได้แก่ เทคนิค Forward Selection เทคนิค Backward Elimination และเทคนิค Evolutionary Selection จากนั้นคัดเลือกคุณลักษณะสำคัญแบบพลวัต (dynamic feature selection) เป็นการคัดเลือกคุณลักษณะที่แปรผันไปตามเป้าหมาย เพื่อให้ได้คุณลักษณะสำคัญที่เหมาะสมที่สุดสำหรับนำไปสร้างแบบจำลอง และสร้างแบบจำลองด้วยเทคนิคต้นไม้ตัดสินใจ (decision tree) ซึ่งเป็นเทคนิคที่เหมาะสมกับการจำแนกประเภทข้อมูล เพื่อการทำนายการได้รับทุน และไม่ได้รับทุนแต่ละประเภท เพื่อให้ค่าความถูกต้อง ค่าความแม่นยำ และค่าประสิทธิภาพโดยรวมสูง ซึ่งผลการทำงานที่ได้จะถูก

นำไปใช้ร่วมกับการพิจารณาของคณะกรรมการในการตัดสินใจให้ทุนแก่นักศึกษาของวิทยาลัยฯ อย่างยุติธรรม และโปร่งใส

2. วัตถุประสงค์การวิจัย

2.1 เพื่อบูรณาการอัลกอริทึมการเลือกคุณลักษณะสำคัญแบบพลวัตกับการจำแนกต้นไม้ตัดสินใจบนพื้นฐานเงื่อนไขหลายเป้าหมาย

2.2 เพื่อประเมินประสิทธิภาพการเลือกคุณลักษณะสำคัญแบบพลวัตสำหรับความเหมาะสมที่สุดของการจำแนกต้นไม้ตัดสินใจบนพื้นฐานเงื่อนไขหลายเป้าหมาย

3. ขอบเขตของการวิจัย

3.1 ข้อมูลนักศึกษาวิทยาลัยเทคโนโลยีทางการแพทย์และสาธารณสุข กาญจนาภิเษก ประกอบด้วยข้อมูลคุณลักษณะสำคัญจำนวน 29 คุณลักษณะสำคัญ แถวข้อมูลจำนวน 500 แถว

3.2 วิเคราะห์คุณลักษณะสำคัญ จำนวน 29 คุณลักษณะสำคัญ แถวข้อมูลจำนวน 500 แถว และกำหนดเงื่อนไขหลายเป้าหมาย จำนวน 3 เป้าหมาย ได้แก่ 1) เป้าหมาย A คือ ทุนได้เปล่า 2) เป้าหมาย B คือ ทุนกู้ยืมเพื่อการศึกษา และ 3) เป้าหมาย C คือ ทุนขาดแคลนทุนทรัพย์

3.3 วิเคราะห์เทคนิคการคัดเลือกคุณลักษณะที่สำคัญทั้งหมด 7 เทคนิคของการเรียนรู้ของเครื่อง (Machine Learning : ML) เลือกคุณลักษณะสำคัญแบบพลวัต และสร้างแบบจำลอง (model) การจำแนกต้นไม้ตัดสินใจ (decision tree classification)

3.4 ประเมินประสิทธิภาพ ความถูกต้อง ความแม่นยำ ค่าเรียกคืน และค่าประสิทธิภาพโดยรวมของการเลือกคุณลักษณะสำคัญแบบพลวัตสำหรับความเหมาะสมที่สุดของการจำแนกต้นไม้ตัดสินใจบนพื้นฐานเงื่อนไขหลายเป้าหมาย

4. นิยามศัพท์เฉพาะ

4.1 คุณลักษณะ (feature) หมายถึง ข้อมูลคุณสมบัติเฉพาะที่เกี่ยวข้องกับชุดข้อมูลที่จะนำมาสร้างอัลกอริทึม

4.2 การคัดเลือกคุณลักษณะ (feature selection) หมายถึง การคัดเลือกคุณลักษณะ (feature selection) เป็นกระบวนการตัดคุณลักษณะที่ไม่มีความสำคัญออก เหลือเพียงคุณลักษณะสำคัญที่เหมาะสำหรับการสร้างแบบจำลองเท่านั้น ซึ่งทำให้การประมวลผลรวดเร็ว มีประสิทธิภาพและมีค่าความถูกต้องสูงขึ้น การคัดเลือกคุณลักษณะเป็นเทคนิคที่ช่วยลดจำนวนตัวแปรที่ใช้สำหรับสร้างตัวแบบพยากรณ์ เพื่อเลือกตัวแปรที่ดีที่สุดหรือเป็นตัวแปรที่สำคัญ โดยเทคนิคการคัดเลือกคุณลักษณะ แบ่งเป็น 3 ประเภท ได้แก่ แบบกรอง (Filter method) แบบควบรวม (Wrapper method) และแบบฝังตัว (Embedded method)

4.3 วิธีแบบกรอง (filter approach) หมายถึง เป็นวิธีที่คัดเลือกคุณลักษณะสำคัญ โดยการคำนวณค่าน้ำหนักหรือค่าความสัมพันธ์ของแต่ละคุณลักษณะ งานวิจัยนี้เลือกใช้เทคนิค Correlation base Feature Selection เทคนิค Information Gain เทคนิค Gain Ratio และเทคนิค Chi Square

4.4 วิธีแบบควบรวม (wrapper approach) หมายถึง เป็นวิธีคัดเลือกคุณลักษณะสำคัญโดยการคำนวณค่าน้ำหนักการวัดค่าความถูกต้องในการแบ่งกลุ่มข้อมูล มาสร้างเซตคุณลักษณะใหม่โดยการเพิ่มหรือลดจำนวนคุณลักษณะจากเซตเดิม งานวิจัยนี้เลือกใช้เทคนิค Forward Selection เทคนิค Backward Elimination และเทคนิค Evolutionary Selection

4.5 การเลือกคุณลักษณะสำคัญแบบพลวัต (dynamic feature selection) หมายถึง เป็นวิธีการคัดเลือกคุณลักษณะสำคัญที่แปรผันไปตามเงื่อนไขหลายเป้าหมาย (multi-target conditions) โดยได้เปรียบเทียบเทคนิคการคัดเลือกคุณลักษณะทั้งหมด 7 เทคนิค โดยเลือกเทคนิคที่สามารถลดจำนวนคุณลักษณะของข้อมูลลง และสามารถเพิ่มความแม่นยำในการวัดค่าความถูกต้องของตัวแบบการพยากรณ์ ด้วยงานวิจัยนี้กำหนดเป้าหมาย จำนวน 3 เป้าหมาย ซึ่งแต่ละเป้าหมายไม่มีความสัมพันธ์กันหรือเป็นอิสระต่อกัน (independent) ผู้วิจัยจึงทำการเลือกคุณลักษณะสำคัญแบบพลวัต ซึ่งคุณลักษณะสำคัญที่ได้จะแปรผันไปตามเป้าหมาย โดยเลือกจากเทคนิคการคัดเลือกคุณลักษณะที่หลากหลาย ทั้งหมด 7 เทคนิค เพื่อให้ได้คุณลักษณะสำคัญที่ความเหมาะสมกับวิธีการจำแนกต้นไม้ตัดสินใจ

4.6 เป้าหมาย (target) หมายถึง ผลลัพธ์ที่ต้องการพยากรณ์ โดยงานวิจัยนี้เป็นการพยากรณ์การได้รับทุนการศึกษา และไม่ได้รับทุนการศึกษาประเภทต่างๆ

4.7 เงื่อนไขหลายเป้าหมาย (multi-target) หมายถึง การกำหนดผลลัพธ์ที่ต้องการพยากรณ์ตั้งแต่ 2 เป้าหมายขึ้นไป ซึ่งงานวิจัยนี้มีเป้าหมาย จำนวน 3 เป้าหมาย ได้แก่ 1) ทุนได้เปล่า 2) ทุนกู้ยืมเพื่อการศึกษา และ 3) ทุนขาดแคลนทุนทรัพย์

4.8 การจำแนกประเภท (classification) หมายถึง เป็นเทคนิคการเรียนรู้ของเครื่องที่ใช้ในการจำแนกประเภทข้อมูล งานวิจัยนี้กำหนดการจำแนกประเภทข้อมูลของแต่ละเป้าหมายแบบไบนารี (binary classification) เป็นการจำแนกประเภทข้อมูลเพียงสองคลาส ได้แก่ ได้รับทุนการศึกษา และไม่ได้รับทุนการศึกษา

4.9 แบบจำลอง (model) หมายถึง รูปแบบนำเสนอที่ถูกสร้างขึ้นแทนข้อเท็จจริงต่าง ๆ โดยอาศัยข้อมูลที่ได้จากการทดลอง

4.10 ต้นไม้ตัดสินใจ (decision tree) หมายถึง เทคนิคการจำแนกประเภทข้อมูลอย่างหนึ่ง มีลักษณะเหมือนโครงสร้างต้นไม้ ที่นิยมค่อนข้างแพร่หลาย เนื่องจากมีความง่ายต่อการทำความเข้าใจผลลัพธ์ และง่ายต่อการนำไปปรับเปลี่ยนให้เป็นกฎการจำแนก (classification rules)

ซึ่งประกอบด้วยโหนดแรกสุด (root node) จากรูทโหนดจะแตกออกเป็นโหนดลูก (child node) โหนดลูกอาจมีลูกของตัวเอง จนถึงโหนดในระดับสุดท้ายที่เรียกว่าโหนดใบ (leaf node)

5. ประโยชน์ที่คาดว่าจะได้รับ

ประโยชน์ที่คาดว่าจะได้รับจากการวิจัยในครั้งนี้ สามารถสรุปได้ดังนี้

5.1 ได้อัลกอริทึมการเลือกคุณลักษณะสำคัญแบบพลวัตกับการจำแนกต้นไม้ตัดสินใจบนพื้นฐานเงื่อนไขหลายเป้าหมาย สำหรับเลือกคุณลักษณะสำคัญของข้อมูลให้สอดคล้องกับผลลัพธ์หรือเป้าหมายที่ต้องการ

5.2 ได้อัลกอริทึมการเลือกคุณลักษณะสำคัญแบบพลวัตกับการจำแนกต้นไม้ตัดสินใจบนพื้นฐานเงื่อนไขหลายเป้าหมาย ที่มีประสิทธิภาพสูง และสามารถนำไปใช้งานได้จริง

5.3 สามารถนำไปใช้งานร่วมกับการพิจารณาทุนของวิทยาลัยเทคโนโลยีทางการแพทย์และสาธารณสุข กาญจนาภิเษก เพื่อส่งเสริมการให้ทุนอย่างถูกต้อง โปร่งใส ไม่เอนเอียง และเป็นที่น่าเชื่อถือ

บทที่ 2
วรรณกรรมที่เกี่ยวข้อง

งานวิจัยเรื่อง การเลือกคุณลักษณะสำคัญแบบพลวัตสำหรับความเหมาะสมที่สุดของการจำแนกต้นไม้ตัดสินใจบนพื้นฐานเงื่อนไขหลายเป้าหมาย ผู้วิจัยได้ศึกษาแนวคิดและทฤษฎี รวมทั้งงานวิจัยที่เกี่ยวข้องเพื่อนำมาประกอบการวิจัย ดังนี้

1. แนวคิดและทฤษฎี งานวิจัยนี้อาศัยหลักการแนวคิดและทฤษฎีที่เกี่ยวข้องทั้งหมด 8 แนวคิดและทฤษฎี ได้แก่
    - 1.1 การเรียนรู้ของเครื่อง (Machine Learning : ML)
    - 1.2 การคัดเลือกคุณลักษณะสำคัญ (feature selection)
    - 1.3 การคัดเลือกคุณลักษณะสำคัญแบบพลวัต (dynamic feature selection)
    - 1.4 การจัดการความไม่สมดุลของข้อมูล (imbalanced data classification)
    - 1.5 การสุ่มตัวอย่างกลุ่มน้อยสังเคราะห์ (Synthetic Minority Oversampling Technique: SMOTE)
    - 1.6 เทคนิคการจำแนกประเภทข้อมูล (classification)
    - 1.7 เทคนิคต้นไม้ตัดสินใจ (decision tree)
    - 1.8 การวัดประสิทธิภาพแบบจำลอง (evaluation)
2. งานวิจัยที่เกี่ยวข้อง งานวิจัยนี้อ้างอิงงานวิจัยที่เกี่ยวข้องทั้งหมด 10 งานวิจัย ได้แก่
    - 2.1 งานวิจัย ก งานวิจัยของนิภาพร ชนะมาร และพรรณี สิทธิเดช (2557)
    - 2.2 งานวิจัย ข งานวิจัยของภรัณยา ปาลวิสุทธิ์ (2559)
    - 2.3 งานวิจัย ค งานวิจัยของรัชพล กลัดชื่น และจรัญ แสนราช (2561)
    - 2.4 งานวิจัย ง งานวิจัยของพุทธิพร ธนธรรมเมธี และเยาวเรศ ศิริสถิตกุล (2561)
    - 2.5 งานวิจัย จ งานวิจัยของกาญจน์ ณ ศรีธะ และคณะ (2561)
    - 2.6 งานวิจัย ฉ งานวิจัยของอัจจิมา มณฑาพันธุ์ (2562)
    - 2.7 งานวิจัย ช ฃงานวิจัยของวิชญ์วิสิฐ เกษรสิทธิ์ และคณะ (2563)
    - 2.8 งานวิจัย ซ งานวิจัยของวรการ ใจดี และนพณัฐ วรรณภีร์ (2563)
    - 2.9 งานวิจัย ฌ งานวิจัยของ khan et al. (2011)
    - 2.10 งานวิจัย ญ งานวิจัยของ osiris villacampa (2015)

1. แนวคิดและทฤษฎี

1.1 การเรียนรู้ของเครื่อง (Machine Learning : ML)

การเรียนรู้ของเครื่อง (Machine Learning : ML) เป็นกระบวนการค้นหารูปแบบหรือความสัมพันธ์ที่มีประโยชน์จากข้อมูลที่เก็บรวบรวมมาในอดีต โดยใช้โมเดลต่าง ๆ เพื่อนำมาใช้ในการทำนายการเกิดเหตุการณ์ในอนาคต

1.1.1 ประเภทการเรียนรู้ของเครื่อง

การเรียนรู้ของเครื่อง (Machine Learning : ML) อาศัยอัลกอริทึมที่สามารถแบ่งออกเป็น 3 ประเภท (Prajapati, 2013) ประกอบด้วย (1) การเรียนรู้แบบมีผู้สอน (supervised learning) (2) การเรียนรู้แบบไม่มีผู้สอน (unsupervised learning) และ (3) การเรียนรู้แบบเสริมแรง (reinforcement learning) (ปราโมทย์ ลือนาม, 2561) ดังนี้

1) เทคนิคการเรียนรู้แบบมีผู้สอน (supervised learning) คือ การสร้างโมเดลที่ใช้ในการทำนาย (predictive model) โดยใช้ทั้งข้อมูลนำเข้า (input data) และข้อมูลผลลัพธ์ (output data) ในการสร้างโมเดล การเรียนรู้วิธีนี้เปรียบได้กับการเรียนรู้โดยมีครูฝึกทำหน้าที่ให้คำเฉลย นั่นคือจะต้องมีข้อมูลฝึกสอน (training data) ที่มีผลเฉลย หรือคลาสคำตอบพร้อมกับลาเบล (label) หลังจากทำการเรียนรู้เสร็จสิ้นจะต้องทำการทดสอบโมเดลโดยอาศัยข้อมูลทดสอบ (testing data) เพื่อวัดประสิทธิภาพของโมเดลที่เป็นผลลัพธ์ที่ได้ ตัวอย่างตัวแบบที่สร้างด้วยเทคนิคการเรียนรู้แบบมีผู้สอน (supervised learning) เช่น

1.1) การจำแนก (classification model) เป็นกระบวนการสร้างโมเดลจัดการข้อมูลให้อยู่ในกลุ่มที่กำหนดมาให้จากกลุ่มตัวอย่างข้อมูลที่เรียกว่า ข้อมูลฝึกสอน (training data) แต่ละแถวของข้อมูลประกอบด้วย แอตทริบิวต์ (attribute) จำนวนมาก แอตทริบิวต์นี้อาจจะเป็นค่าต่อเนื่อง (continuous) ที่เก็บข้อมูลประเภทตัวเลข (numeric data) หรือข้อมูลจำแนกประเภท (categorical data) โดยจะมีแอตทริบิวต์ที่ทำหน้าที่ในการจำแนก (classifying attribute) ซึ่งเป็นตัวกำหนดคลาส (class) ของข้อมูล จุดประสงค์ของการจำแนกประเภทข้อมูลคือการสร้างโมเดลเพื่อแยกข้อมูลโดยอาศัยโมเดลที่ถูกสร้างจากข้อมูลอื่น ทำให้สามารถพิจารณาคลาสในข้อมูลที่ยังไม่ได้แบ่งกลุ่มในอนาคตได้

1.2) การถดถอย (regression) ซึ่งข้อแตกต่างระหว่างการถดถอยกับการจำแนกนั้น คือ การถดถอยจะใช้กับข้อมูลประเภทต่อเนื่อง (continuous) ในการสร้างโมเดล ตัวอย่างการถดถอย เช่น โมเดลการทำนายปริมาณน้ำฝนที่จะตก ในรูปแบบของความสัมพันธ์กับตัวแปรต่างๆ เช่น อุณหภูมิ ความชื้นสัมพัทธ์ เป็นต้น

2) เทคนิคการเรียนรู้แบบไม่มีผู้สอน (unsupervised learning) คือ การจัดกลุ่มและการแปลผลข้อมูลโดยอาศัยข้อมูลเข้า (input data) เพียงอย่างเดียว การเรียนรู้นี้ต่างจากการเรียนรู้แบบมีผู้สอน คือ จะไม่มีการระบุผลที่ต้องการ (target) ไว้ก่อน การเรียนรู้นี้ไม่ได้มีลาเบลกำกับเพื่อบอกว่าข้อมูลนั้นคืออะไร แต่การเรียนรู้นี้จะจัดข้อมูลนำเข้า (input) จัดกลุ่ม (cluster) บนพื้นฐานความเหมือน (similarities) และความแตกต่าง (differences) ระหว่างรูปแบบของการนำเข้าข้อมูล (input patterns) เป็นเทคนิคที่พิจารณาข้อมูลเป็นหลัก เช่น พิจารณาว่าข้อมูลมีความสัมพันธ์กันในลักษณะใดบ้าง เทคนิคในประเภทนี้จะแบ่งย่อยได้อีก คือ เทคนิคการค้นหากฎความสัมพันธ์ (association rule) และการแบ่งกลุ่มข้อมูล (clustering) เป็นต้น ตัวอย่างตัวแบบที่สร้างด้วยเทคนิคการเรียนรู้แบบไม่มีผู้สอน (unsupervised learning) เช่น

2.1) การลดมิติข้อมูล (dimensionality reduction) เป็นวิธีที่ถูกนำเสนอเพื่อแก้ไขปัญหาของมิติข้อมูล (curse of dimension) ซึ่งได้รับผลกระทบจากข้อมูลที่ลักษณะมีมิติสูง (high dimension) ที่ไม่มีการจัดการข้อมูลก่อนในเบื้องต้น ข้อมูลที่มีจำนวนมากนั้นมีลักษณะข้อมูลที่กระจัดกระจาย และบางข้อมูลไม่มีประโยชน์ในการวิเคราะห์ ซึ่งจะส่งผลกระทบต่อความถูกต้องของการประมวลผล และบางอัลกอริทึมของประมวลผลไม่สามารถรองรับการทำงานของข้อมูลหรือตัวแปรจำนวนที่มาก ๆ ได้ นอกจากนี้ยังทำให้สิ้นเปลืองทรัพยากรในการประมวลผลอีกด้วย เช่น เวลาในการประมวลผลนาน ใช้หน่วยความจำมากในประมวลผลแต่ละครั้ง

2.2) การแบ่งกลุ่ม (clustering) เป็นการแบ่งข้อมูลออกเป็นกลุ่ม โดยข้อมูลที่อยู่ในกลุ่มเดียวกัน จะมีลักษณะที่คล้ายคลึงกันและข้อมูลที่อยู่ต่างกลุ่มจะมีลักษณะที่แตกต่างกัน การจัดกลุ่มข้อมูลต้องอาศัยการวัดความคล้ายคลึงหรือความแตกต่างระหว่างข้อมูล 2 ตัว ซึ่งสามารถทำได้หลายแบบขึ้นอยู่กับชนิดของข้อมูลและการกำหนดความคล้ายคลึงของข้อมูล ตามวัตถุประสงค์ในการจัดกลุ่มข้อมูล ตัวอย่างการจัดกลุ่มข้อมูล เช่น การจัดกลุ่มลูกค้า การจัดกลุ่มของเอกสารที่มีเนื้อหาคล้ายคลึง เป็นต้น

2.3) การสร้างกฎความสัมพันธ์ (association rules) เป็นการหากฎที่บอกถึงความสัมพันธ์ระหว่างข้อมูลที่มักเกิดขึ้นพร้อมกันอยู่เสมอ โดยศึกษาความสัมพันธ์ระหว่างคุณสมบัติต่าง ๆ ในข้อมูล ผลการวิเคราะห์คือ กลุ่มของกฎความสัมพันธ์ในลักษณะถ้า - แล้ว และมีมาตรวัดคุณภาพของกฎกำกับ ยกตัวอย่างเช่น ข้อมูลรายการที่มีการซื้อขายในแต่ละครั้งหรือเหตุการณ์ที่เกิดขึ้นพร้อมกันภายในรายการ (transection) กฎความสัมพันธ์ที่สร้างได้จะระบุถึงความสัมพันธ์ว่าเมื่อพบเหตุการณ์หนึ่งหรือหลายเหตุการณ์เกิดขึ้น จะมีโอกาสสูงที่เหตุการณ์อีกอย่างหนึ่งหรือหลายเหตุการณ์จะเกิดขึ้นด้วย

3) การเรียนรู้แบบเสริมแรง (reinforcement learning) คือ การเรียนรู้ที่คอมพิวเตอร์ให้ความสนใจต่อเหตุการณ์ที่เกิดในสภาพแวดล้อม ผู้เรียนหรือเอเจนต์ (agent) จะอาศัยการฝังสังเกต (observe) จนเมื่อเหตุการณ์ที่จะต้องตัดสินใจเลือกการกระทำตามข้อกำหนดของนโยบาย (policy) ที่ได้ตั้งไว้ หลังจากได้ดำเนินการกระทำ (action) แล้วผู้เรียนจะได้รับผลตอบแทนเป็นรางวัล (reward) หรือการลงโทษ (penalty) ผลตอบแทนดังกล่าวจะถูกนำไปใช้ในการปรับปรุงนโยบาย (update policy) ที่ดีที่สุด ทำให้เกิดกระบวนการเรียนรู้ กระบวนการปรับปรุงจะทำในลักษณะวนซ้ำ (iteration) จนกระทั่งได้นโยบายที่เหมาะสมที่สุด (optimal policy) หรือได้รางวัลรวมที่ความหวัง (maximum sum of expected rewards) สูงที่สุด

1.1.2 การจัดกลุ่มของอัลกอริทึมการเรียนรู้ของเครื่อง

อัลกอริทึมการเรียนรู้ของเครื่องหากนำมาจัดกลุ่มด้วยเกณฑ์วิธีการเรียนรู้แบบมีผู้สอน และวิธีการเรียนรู้แบบไม่มีผู้สอน และนำเข้าข้อมูลที่นำมาใช้สร้างโมเดลที่มีลักษณะเป็นค่าต่อเนื่อง (continuous data) กับข้อมูลจำแนกประเภท (categorical data) จะแสดงในตารางที่ 2.1

ตารางที่ 2.1 ตารางจำแนกประเภทของอัลกอริทึมแบบ Supervised และ Unsupervised learning

	Supervised learning	Unsupervised learning
Continuous	**รีเกรสชัน (regression)** - การวิเคราะห์การถดถอยเชิงเส้น (linear regression) - การวิเคราะห์การถดถอยเชิงเส้นแบบพหุ (multiple linear regression)	**การจัดกลุ่ม (clustering) และการลดมิติข้อมูล (dimensionality reduction)** - K-Means - Principal Component Analysis: Pca - Singular Value Decomposition: SVD
Categorical	**การจำแนก (classification)** - K-nearest neighbors: knn - Naïve bayes - support vector machine - decision tree	**การวิเคราะห์ความสัมพันธ์ (association analysis)** - Apriori - FP-Growth

งานวิจัยนี้เป็นการวิเคราะห์เชิงทำนาย (predictive analytics) โดยการนำข้อมูลที่มีอยู่เดิมในองค์กร มาทำการวิเคราะห์ด้วยวิธีการเรียนรู้ของเครื่อง (Machine Learning : ML) เพื่อทำการวิเคราะห์โอกาสหรือแนวโน้มของนักศึกษาที่ยื่นขอทุนการศึกษามีโอกาสที่จะได้รับทุนการศึกษาหรือไม่ได้รับทุนการศึกษาประเภทต่างๆ ซึ่งวิธีการนี้เป็นการเรียนรู้แบบมีผู้สอน (supervised learning) ซึ่งอาศัยข้อมูลนำเข้า (input data) และข้อมูลผลลัพธ์ (output data) ในการสร้าง

แบบจำลองการจำแนกประเภทข้อมูล (classification model) เนื่องจากลักษณะของข้อมูลที่นำมาจำแนกเป็นข้อมูลการจำแนกประเภท (categorical data) ได้แก่ ได้รับทุนการศึกษา และไม่ได้รับทุนการศึกษา

1.1.3 กระบวนการเรียนรู้ของเครื่อง

ประยุกต์ใช้กระบวนการดำเนินงาน ซึ่งประกอบด้วย 6 ขั้นตอน (Romsaiyud, W., Schnoor, H., & Hasselbring, W., 2019) ได้แก่ 1) การเก็บรวบรวมข้อมูล (data collection) 2) การเตรียมข้อมูล (data preparation) 3) การเลือกโมเดล (choose the model) 4) การฝึกสอนโมเดล (training the model) 5) การวัดประสิทธิภาพโมเดล (evaluation the model) และ 6) การปรับค่าพารามิเตอร์ (parameter tuning) และการใช้งานการทำนายแบบจำลอง (model prediction) มีรายละเอียดขั้นตอน ดังนี้

1) การเก็บรวบรวมข้อมูล (data collection)

เป็นขั้นตอนการเก็บรวบรวมข้อมูลที่เกี่ยวข้อง ที่จะนำมาวิเคราะห์โดยทำความเข้าใจข้อมูลและศึกษาลักษณะความสัมพันธ์ข้อมูล รูปแบบ โดยใช้กระบวนการวิเคราะห์ เทคนิควิเคราะห์ และใช้เทคโนโลยีสารสนเทศให้ถูกต้อง เหมาะสม และสอดคล้องกับข้อมูล เพื่อให้สอดคล้องกับความต้องการในการนำไปใช้ประโยชน์ รวมทั้งประเมินคุณภาพของข้อมูล กำหนดคุณสมบัติข้อมูล และทำการคัดเลือกข้อมูลที่จะนำมาวิเคราะห์ตรวจสอบความสมบูรณ์และความถูกต้องของข้อมูล

2) การเตรียมข้อมูล (data preparation)

เป็นขั้นตอนการเตรียมข้อมูล โดยเป็นขั้นตอนที่ใช้ระยะเวลานานที่สุด เนื่องจากโมเดลที่ได้นั้นจะให้ผลลัพธ์ที่ถูกต้องหรือไม่นั้นขึ้นอยู่กับคุณภาพของข้อมูลที่ใช้ โดยเริ่มต้นจากการเก็บรวมรวมข้อมูลจากแหล่งข้อมูล (data source) จากนั้นนำข้อมูลเข้าสู่ขั้นตอนการเตรียมข้อมูล ซึ่งสามารถแบ่งออกได้เป็น 3 ขั้นตอนย่อย ได้แก่ 1) ขั้นตอนการคัดเลือกข้อมูล (data selection) ซึ่งเป็นการกำหนดเป้าหมายว่าเราจะทำการวิเคราะห์อะไร โดยขั้นตอนนี้จะทำการเลือกเฉพาะข้อมูลที่เกี่ยวข้องกับสิ่งที่จะทำ 2) ขั้นตอนการกลั่นกรองข้อมูล (data cleaning) เป็นการลบข้อมูลที่มีความซ้ำซ้อน และทำการแก้ไขข้อมูลที่ผิดพลาด เช่น รูปแบบของข้อมูลผิด ข้อมูลหายไป และข้อมูล outlier ที่แปลกแยกจากคนอื่น และ 3) ขั้นตอนการแปลงรูปของข้อมูล (data transformation) เป็นขั้นตอนการเตรียมข้อมูลในอยู่ในรูปแบบที่พร้อมจะนำไปใช้วิเคราะห์ด้วยอัลกอริทึมต่าง ๆ และนำข้อมูลเข้าสู่กระบวนการคัดเลือกคุณลักษณะ (feature selection) เพื่อคัดเลือกคุณลักษณะที่สำคัญสำหรับการสร้างโมเดลจำแนกประเภทข้อมูล โดยใช้เทคนิคการคัดเลือกคุณลักษณะต่างๆ เข้ามาช่วยในการคัดเลือก ซึ่งเป็นการตัดคุณลักษณะที่ไม่มีความสำคัญในการสร้างแบบจำลองออก เหลือเพียงคุณลักษณะสำคัญที่เหมาะสำหรับการสร้างแบบจำลองจำแนกประเภทเท่านั้น ซึ่งการลดจำนวนคุณลักษณะลงจะทำให้การประมวลผลรวดเร็วมากขึ้น ทั้งนี้เพื่อให้

แบบจำลองการพยากรณ์บนพื้นฐานเงื่อนไขหลายเป้าหมายให้มีประสิทธิภาพเพิ่มมากขึ้นและมีค่าความถูกต้องสูงขึ้น

3) การเลือกโมเดล (choose the model)

เป็นขั้นตอนการเลือกวิธีการวิเคราะห์ข้อมูล โดยการเลือกอัลกอริทึมที่เหมาะสมที่สุดเพื่อให้ได้ผลลัพธ์ที่ดีที่สุด ในการดำเนินการสร้างโมเดลจะต้องมีการกำหนดรูปแบบที่เหมาะสม นำมาสร้างแบบจำลองตามอัลกอริทึมที่เลือก

4) การฝึกสอนโมเดล (training the model)

เป็นขั้นตอนการนำข้อมูลมาสร้างแบบจำลอง โดยการสร้างแบบจำลองอาจมีการนำเทคนิคหลายเทคนิคมาใช้ร่วมกัน และในขั้นตอนนี้บางกรณีสามารถย้อนกลับไปขั้นตอนที่ 2 เพื่อเตรียมการคัดเลือกตัวอย่างข้อมูลและเพิ่มเติมตัวแปรในการวิเคราะห์ได้

5) การวัดประสิทธิภาพโมเดล (evaluation the model)

เป็นขั้นตอนที่ประเมินผลลัพธ์และเปรียบเทียบประสิทธิผลของโมเดลว่าจะสามารถนำไปใช้ได้หรือไม่ ซึ่งโดยปกติจะเป็นการประเมินรูปแบบ (model) ที่ได้จากขั้นตอนที่ 4 ว่าสามารถให้ผลลัพธ์ที่จะสามารถนำไปใช้งานตามวัตถุประสงค์ของโครงการได้มากน้อยเพียงใด ขั้นตอนนี้อาจจะย้อนกลับไปขั้นตอนที่ 4 เพื่อสร้างรูปแบบใหม่ที่มีความสมบูรณ์มากยิ่งขึ้น

6) การปรับค่าพารามิเตอร์ (parameter tuning)

เป็นขั้นตอนการปรับค่าพารามิเตอร์ที่สำคัญให้เหมาะสม (optimize parameter) เพื่อเพิ่มประสิทธิภาพให้กับแบบจำลอง และหลังจากได้รูปแบบที่สมบูรณ์ จากขั้นตอนที่แล้วจะนำไปใช้งานจริง ซึ่งผลลัพธ์จากการใช้งานสามารถนำมาใช้ในการวางแผนเพื่อเริ่มทำโครงการต่อไปในอนาคต

โดยงานวิจัยนี้จะเน้นการคัดเลือกคุณลักษณะ (feature selection) ซึ่งอยู่ในขั้นตอนที่ 2 การเตรียมข้อมูล (data preparation) เนื่องจากการคัดเลือกคุณลักษณะเป็นสิ่งสำคัญ ซึ่งการดำเนินการในการจัดเก็บข้อมูลงานวิจัย จะต้องคัดเลือกตัวอย่างข้อมูลและตัวแปรที่จะใช้ในการวิเคราะห์มาทำตามขั้นตอนย่อยทั้ง 3 ขั้นตอน ได้แก่ 1) ขั้นตอนการคัดเลือกข้อมูล (data selection) 2) ขั้นตอนการกลั่นกรองข้อมูล (data cleaning) และ 3) ขั้นตอนการแปลงรูปของข้อมูล (data transformation) โดยผู้วิจัยนำข้อมูลที่จัดเก็บจากแหล่งข้อมูลมาผ่านกระบวนการทำความสะอาดข้อมูลและแก้ไขข้อมูลให้สมบูรณ์ถูกต้อง หรืออาจจะต้องลดรูปข้อมูล รวมทั้งกำจัดข้อมูลที่มีลักษณะผิดปกติซึ่งอาจทำให้การวิเคราะห์มีความผิดพลาด จากนั้นทำการแปลงข้อมูลให้อยู่ในรูปแบบที่เหมาะสมที่สามารถนำมาวิเคราะห์ได้ และนำข้อมูลเข้าสู่กระบวนการคัดเลือกคุณลักษณะ (feature selection) เพื่อคัดเลือกคุณลักษณะที่สำคัญสำหรับการสร้างโมเดลจำแนกประเภทข้อมูล โดยใช้เทคนิคการคัดเลือกคุณลักษณะต่างๆ เข้ามาช่วยในการคัดเลือก ซึ่งเป็นการตัดคุณลักษณะที่ไม่มี

ความสำคัญในการสร้างแบบจำลองออก เหลือเพียงคุณลักษณะสำคัญที่เหมาะสำหรับการสร้างแบบจำลองจำแนกประเภทเท่านั้น ซึ่งการลดจำนวนคุณลักษณะลงจะทำให้การประมวลผลรวดเร็วมากขึ้น ทั้งนี้เพื่อให้แบบจำลองการพยากรณ์บนพื้นฐานเงื่อนไขหลายเป้าหมายให้มีประสิทธิภาพเพิ่มมากขึ้นและมีค่าความถูกต้องสูงขึ้น

1.2 การคัดเลือกคุณลักษณะสำคัญ (feature selection)

การคัดเลือกคุณลักษณะ (feature selection) เป็นกระบวนการคัดคุณลักษณะที่ไม่มีความสำคัญหรือไม่เกี่ยวข้องกับเป้าหมายออก เพื่อให้ได้คุณลักษณะสำคัญที่มีความเกี่ยวข้องกับเป้าหมายสูงและเหมาะสำหรับการสร้างแบบจำลอง ซึ่งทำให้การประมวลผลรวดเร็ว และค่าความถูกต้องสูงขึ้น การคัดเลือกคุณลักษณะเป็นเทคนิคที่ช่วยลดจำนวนตัวแปรที่ใช้สำหรับสร้างตัวแบบพยากรณ์ เพื่อเลือกตัวแปรที่ดีที่สุดหรือเป็นตัวแปรที่สำคัญ โดยเทคนิคการคัดเลือกคุณลักษณะแบ่งเป็น 3 ประเภท ได้แก่ แบบกรอง (Filter method) แบบควบรวม (Wrapper method) และแบบฝังตัว (Embedded method) งานวิจัยนี้เลือกใช้วิธีการคัดเลือกคุณลักษณะ 2 วิธี ดังนี้

1.2.1 วิธีแบบกรอง (Filter approach) เป็นวิธีที่คัดเลือกคุณลักษณะสำคัญ โดยการคำนวณค่าน้ำหนักหรือค่าความสัมพันธ์ของแต่ละคุณลักษณะ (อัจจิมา มณฑาพันธุ์, 2560) โดยเลือกเฉพาะคุณลักษณะที่มีความสำคัญเก็บเอาไว้ งานวิจัยนี้เลือกใช้เทคนิค Correlation base Feature Selection เทคนิค Information Gain เทคนิค Gain Ratio และเทคนิค Chi Square เป็นต้น

1) เทคนิค Correlation base Feature Selection

เทคนิค Correlation base Feature Selection เป็นเทคนิคในการเลือกคุณสมบัติของคุณลักษณะโดยใช้การพิจารณาบนพื้นฐานความสัมพันธ์ ของกลุ่มคุณลักษณะที่ได้จากการประเมินค่าจากความสามารถในการคาดการณ์ คุณลักษณะที่คัดเลือกใช้สำหรับจำแนกประเภทข้อมูล และยังสามารถจัดการกับคุณลักษณะที่ไม่เกี่ยวข้อง CFS จะจัดอันดับกลุ่มย่อยของมิติข้อมูล ทำการคัดเลือกกลุ่มย่อยของมิติข้อมูลที่มีความสัมพันธ์กันสูงกับคลาส และไม่มีความสัมพันธ์กับคลาสอื่น ๆ สำหรับมิติข้อมูลที่ไม่เกี่ยวข้องหรือมีความสัมพันธ์ต่ำกับคลาสจะถูกลบทิ้ง มิติข้อมูลที่ซ้ำซ้อนจะถูกขจัดออกไปจากกลุ่มมิติข้อมูลที่มีความสัมพันธ์สูง สมการประเมินกลุ่มย่อยของมิติข้อมูลแบบ CFS แสดงในสมการที่ 1 (ภัทราวุฒิ แสงศิริ, 2553)

$$M_s = \frac{k\overline{r_{cf}}}{\sqrt{k+k(k-1)\overline{r_{ff}}}} \qquad (1)$$

โดยที่ M_s คือ ค่าที่ค้นหาได้ของมิติข้อมูลกลุ่มย่อย s ซึ่งประกอบด้วยมิติข้อมูล k

$\overline{r_{sf}}$ คือ ค่าเฉลี่ยความสัมพันธ์ของตัวแปรกับคลาส $(f \in s)$

$\overline{r_{ff}}$ คือ ค่าเฉลี่ยความสัมพันธ์ระหว่างมิติข้อมูล

2) เทคนิค Information Gain

เทคนิค Information Gain เป็นเทคนิคที่ใช้ในการชี้วัดเพื่อเลือกคุณลักษณะข้อมูลที่ต้องการที่น้อยที่สุดที่เหมาะสมในการนำไปใช้ระบุ โดยคำนวณหาค่า gain สำหรับแต่ละมิติข้อมูล ซึ่งข้อมูลใดมีค่า gain สูงสุด จะถูกคัดเลือกเพื่อนำมาใช้ระบุ สมการที่ 2 แสดงการคำนวณค่า Information Gain หรือค่า entropy ของชุดข้อมูลทั้งหมด สมการที่ 3 แสดงการคำนวณค่า entropy ของชุดมิติข้อมูลในแต่ละคุณลักษณะ สมการที่ 4 การคำนวณหาค่า Information Gain สำหรับการพิจารณามิติของข้อมูลคุณลักษณะ A (อัจจิมา มณฑาพันธุ์, 2560)

$$E(D) = -\sum_{i=1}^{n} p_i \log_2(p_i) \quad (2)$$

$$E_A(D) = \sum_{j=1}^{m} \frac{|D_j|}{D} x \; E(D_j) \quad (3)$$

$$Gain(A) = E(D) - E_A(D) \quad (4)$$

โดยที่ P_i คือ ค่าความน่าจะเป็นที่เรคอร์ดหนึ่ง ๆ จะมีหมวดหมู่ของข้อมูล

หรือกล่าวว่า การคำนวณหาค่า Information Gain คือการวัดค่า entropy ก่อนที่จะมีการแบ่งข้อมูลออกตามมิติข้อมูลและหลังการแบ่งว่ามีประสิทธิภาพดีขึ้นหรือไม่ ถ้ามีประสิทธิภาพดีขึ้นค่า Information Gain จะมีค่าสูง

3) เทคนิค Gain Ratio

เทคนิค Gain Ratio เป็นตัวชี้วัดการแบ่งชุดข้อมูลออกเป็นชุดข้อมูลย่อยที่พัฒนามาจาก Information Gain เนื่องจากการใช้ Information Gain ในการแบ่งชุดข้อมูลจะมีโอกาสทำให้เกิดความเอนเอียงขึ้น เมื่อคุณลักษณะที่ทำการพิจารณาได้ค่า gain ที่สูงเป็นจำนวนมาก ทำให้คุณลักษณะที่ถูกคัดเลือกไม่ถูกต้อง ตัวอย่างเช่น พิจารณาคุณลักษณะที่ทำหน้าที่เป็นตัวระบุเฉพาะ เช่น รหัสผลิตภัณฑ์ การแยกรหัสผลิตภัณฑ์ จะส่งผลให้มีชุดข้อมูลย่อยจำนวนมาก แต่ละชุดข้อมูลย่อยอมีเพียงหนึ่ง record ซึ่งเมื่อนำมาหาค่า Information Gain จะได้ค่าที่สูงจำนวนมากนั่นเอง (อัจจิมา มณฑาพันธุ์, 2560)

จากความเอนเอียง ทำให้มีการพัฒนาตัวชี้วักการแบ่งข้อมูลใหม่ที่เรียกว่า Gain Ratio โดยประยุกต์ใช้การทำนอร์มัลไลซ์ค่า Information Gain ด้วยการใช้ค่า “split information” ซึ่งสามารถคำนวณได้ตามสมการที่ 5 (อัจจิมา มณฑาพันธุ์, 2560)

$$SplitInfo_A(D) = \sum_{j=1}^{m} \frac{|D_j|}{|D|} x \; \log_2 \frac{|D_j|}{|D|} \quad (5)$$

โดยค่า $SplitInfo_A(D)$ หมายถึง ปริมาณข้อมูลที่ถูกพิจารณาโดยการแบ่งข้อมูลในชุดข้อมูลในชุดข้อมูล D ออกเป็น m ชุดข้อมูลย่อยตามค่าคุณลักษณะ A โดยหลังจากทำการคำนวณหาค่า $SplitInfo_A(D)$ แล้วเราจะสามารถคำนวณหาค่า Gain Ratio ได้ดังสมการที่ 6 (อัจจิมา มณฑาพันธุ์, 2560)

$$GainRatio\ (A) = \frac{Gain(A)}{SplitInfo(A)} \quad (6)$$

4) เทคนิค Chi Square

เทคนิค Chi Square (x^2) การประเมินค่าของคุณลักษณะโดยใช้การคำนวณค่า chi-square ทางสถิติ เพื่อศึกษาว่าการแจกแจงความถี่ของตัวแปรคุณลักษณะเป็นไปตามรูปแบบที่กำหนดไว้หรือไม่ ดังแสดงในสมการที่ 7 (อัจจิมา มณฑาพันธุ์, 2560)

$$x^2 = \sum_{i=1}^{n} \frac{(O_i - E_j)^2}{E_i} \quad (7)$$

โดย $O1, O2, \dots On$ เป็นความถี่ของตัวแปรที่ได้จากการศึกษา

$E1, E2, \dots En$ เป็นความถี่ที่คาดหวัง (หรือความถี่ที่ควรจะเป็น)

1.2.2 วิธีแบบควบรวม (Wrapper approach) เป็นวิธีคัดเลือกคุณลักษณะสำคัญโดยการคำนวณค่าน้ำหนักการวัดค่าความถูกต้องในการแบ่งกลุ่มข้อมูล มาสร้างเซตคุณลักษณะใหม่โดยการเพิ่มหรือลดจำนวนคุณลักษณะจากเซตเดิม (อัจจิมา มณฑาพันธุ์, 2560) งานวิจัยนี้เลือกใช้เทคนิค Forward Selection เทคนิค Backward Elimination และเทคนิค Evolutionary Selection เป็นต้น

1) เทคนิค Forward Selection

เทคนิค Forward Selection เป็นเทคนิคที่ใช้วิธีการคัดเลือกคุณลักษณะซึ่งเป็นตัวแปรอิสระเข้ามาในสมการทีละตัว ขั้นแรกพิจารณาจากค่าสัมบูรณ์ของค่าสัมประสิทธิ์สหสัมพันธ์ (correlation coefficient) ระหว่างตัวแปรตามกับตัวแปรอิสระที่ให้ค่าสูงสุด หากตัวแปรอิสระตัวนั้นมีคุณสมบัติตามเกณฑ์นำเข้า ก็จะถูกนำเข้าสมการ หลังจากนั้นเป็นการพิจารณาสัมประสิทธิ์สหสัมพันธ์บางส่วนระหว่างตัวแปรตามกับตัวแปรอิสระที่ไม่อยู่ในสมการถดถอยทีละตัว ตัวแปรอิสระที่มีค่าสัมบูรณ์ของค่าสัมประสิทธิ์สหสัมพันธ์บางส่วนสูงสุดจะถูกนำเข้าสมการ หากตัวแปรนั้นมีคุณสมบัติตามเกณฑ์นำเข้า ขั้นตอนจะทำซ้ำจนกระทั่งพบว่าไม่สามารถนำตัวแปรอิสระเข้าสมการได้จึงหยุด

2) เทคนิค Backward Elimination

เทคนิค Backward Elimination เป็นเทคนิคที่พยายามคัดเลือกตัวแปรที่ดีสุดและได้โมเดลประหยัดในการพยากรณ์เช่นเดียวกัน โดยตอนแรกจะนำตัวแปรพยากรณ์ทุกตัวเข้ามาในสมการและดำเนินการพิจารณาตัวแปรพยากรณ์ที่มีค่าสัมประสิทธิ์สหสัมพันธ์บางส่วน (partial correlation) กับตัวแปรเกณฑ์ และควบคุมอิทธิพลของตัวแปรพยากรณ์อื่น ๆ ซึ่งมีค่าต่ำที่สุดออกจากสมการ แล้วจึงดำเนินการทดสอบว่า ค่า r2 ลดลงอย่างมีนัยสำคัญทางสถิติหรือไม่ หากพบว่า ลดลงอย่างไม่มีนัยสำคัญทางสถิติแสดงว่าตัวแปรดังกล่าวไม่ได้มีส่วนทำให้การพยากรณ์ตัวแปรเกณฑ์เพิ่มขึ้นเลย แสดงว่าสามารถขจัดออกจากสมการได้ จากนั้นจึงดำเนินการขจัดตัวแปรพยากรณ์ที่มีความสำคัญน้อยรองลงมาออกไปอีก โดยใช้วิธีพิจารณาเช่นเดียวกันซึ่งการขจัดตัวแปรพยากรณ์จะสิ้นสุดเมื่อพบว่า มีผลทำให้ค่า r2 ลดลงอย่างมีนัยสำคัญทางสถิติ หมายความว่า ตัวแปรดังกล่าวมีความสำคัญต่อการพยากรณ์ตัวแปรตาม หากขจัดตัวแปรดังกล่าวออกจากสมการจะทำให้อำนาจการพยากรณ์ตัวแปรเกณฑ์ลดลงจึงต้องคงตัวแปรพยากรณ์ดังกล่าวไว้ในสมการพยากรณ์ต่อไป (ทรงศักดิ์ ภูสีอ่อน, 2554 : 284 ; Brian S. Everitt. 2010: 93)

เทคนิค Backward Elimination คือ การนำตัวแปรอิสระทั้งหมดเข้าสู่สมการ หลังจากนั้นจะคัดเลือกตัวแปรอิสระที่มีความสัมพันธ์กับตัวแปรตามน้อยที่สุดออกจากสมการดำเนินการซ้ำจนไม่สามารถคัดเลือกตัวแปรอิสระออกจากสมการถดถอยได้ ก็จะเหลือสมการถดถอยที่ตัวแปรอิสระมีความสัมพันธ์กับตัวแปรตามเท่านั้น

3) เทคนิค Evolutionary Selection

เทคนิค Evolutionary Selection เป็นเทคนิคการคัดเลือกคุณลักษณะโดยนำข้อมูลคุณลักษณะมาหาค่าประสิทธิภาพในการคาดการณ์คำตอบ จากนั้นจึงนำคุณลักษณะมาจับคู่กัน แล้วหาค่าประสิทธิภาพอีกครั้ง หากค่าประสิทธิภาพในการคาดการณ์สูงขึ้นจะเก็บข้อมูลนั้นไว้ แล้วทำการเลือกคุณลักษณะอื่นเข้าไปเพิ่ม หากค่าประสิทธิภาพต่ำลงจะถอดคุณลักษณะนั้นออก แล้วเลือกคุณลักษณะอื่นเข้าไป ดังสมการที่ 8 (อัจจิมา มณฑาพันธุ์, 2560)

$$IG(parent\ child) = Entropy(parent) - [p(c_1)\ x\ Entropy(c_1) + p(c_2) x Entropy(c_2) \ldots] \quad (8)$$

โดยที่ $Entropy(c_1)$ คือ $-p(c_1)\log p(c_1)$

$p(c_1)$ คือ ค่าความน่าจะเป็นของค่า c_1

c คือ ปัจจัยคุณลักษณะแต่ละตัวที่เกี่ยวข้อง

ทั้งนี้ค่า entropy จะใช้ในการวัดค่าความแตกต่างกันของข้อมูล ซึ่งถ้าข้อมูลมีความแตกต่างกันน้อยจะมีค่า entropy ต่ำ และถ้าข้อมูลมีความแตกต่างกันมากจะมีค่า entropy สูง (เอกสิทธิ์ พัชรวงศ์ศักดา, 2557)

งานวิจัยนี้ผู้วิจัยเลือกใช้วิธีการคัดเลือกคุณลักษณะ (feature selection) เพื่อคัดเลือกคุณลักษณะข้อมูลซึ่งเป็นปัจจัยสำคัญที่มีความสัมพันธ์กับเป้าหมายที่ต้องการ ในการคัดเลือกคุณสมบัติของชุดข้อมูลที่มีจำนวนมิติสูงหรือมีความหลากหลายของคุณลักษณะข้อมูลมาก ให้มีจำนวนคุณลักษณะข้อมูลที่เหมาะสมกับแบบจำลอง และช่วยเพิ่มประสิทธิภาพความถูกต้องให้กับแบบจำลองมากที่สุด โดยงานวิจัยนี้ทำการเปรียบเทียบเทคนิคทั้งหมด 7 เทคนิคของการเรียนรู้ของเครื่อง (Machine Learning : ML) ได้แก่ เทคนิค Correlation base Feature Selection เทคนิค Information Gain เทคนิค Gain Ratio เทคนิค Chi Square เทคนิค Forward Selection เทคนิค Backward Elimination และเทคนิค Evolutionary Selection เนื่องจากเทคนิคทั้ง 7 เทคนิค เป็นเทคนิควิธีที่นิยมกันอย่างแพร่หลายในงานวิจัยอื่น ๆ มากมาย และมีการประมวลผลรวดเร็ว สามารถตีความจากค่าน้ำหนัก (weight) ของแต่ละคุณลักษณะ ซึ่งเทคนิคดังกล่าวจะช่วยลดมิติของข้อมูลลงและเพิ่มประสิทธิภาพแบบจำลอง โดยในแต่ละเทคนิควิธีจะมีความเหมาะสมกับข้อมูลที่แตกต่างกันออกไป

1.3 การคัดเลือกคุณลักษณะสำคัญแบบพลวัต (dynamic feature selection)

การเลือกคุณลักษณะสำคัญแบบพลวัต (dynamic feature selection) เป็นวิธีการคัดเลือกคุณลักษณะสำคัญที่แปรผันไปตามเงื่อนไขหลายเป้าหมาย (multi-target conditions) โดยวิเคราะห์จากความสัมพันธ์ของเป้าหมาย การคัดเลือกคุณลักษณะสำคัญแบบพลวัต เป็นวิธีการที่เหมาะสมกับเป้าหมายที่ไม่มีความสัมพันธ์กันหรือเป็นอิสระต่อกัน (independent) เพื่อใช้ในการคัดเลือกหรือลดจำนวนคุณลักษณะลง เหลือเพียงคุณลักษณะสำคัญที่สอดคล้องกับเป้าหมาย

ซึ่งในการกำหนดคลาสเป้าหมายแบบเดิมจะกำหนดในลักษณะของ single-label ใช้สำหรับวิเคราะห์ข้อมูลเพียงเป้าหมายเดียว โดยการจำแนกประเภทข้อมูลแบบไบนารี (binary) เช่น “ใช่” หรือ “ไม่ใช่” ดังภาพที่ 2.1 หรือการกำหนดคลาสเป้าหมายแบบหลายค่าในรูปแบบของ single label ดังภาพที่ 2.2 และการแก้ปัญหาการจำแนกประเภทข้อมูลแบบหลายลาเบล โดยวิธีการกำหนดคลาสเป้าหมายแบบไบนารี (binary) ในรูปแบบ multi-label ดังภาพที่ 2.3

single label	
instance	class
1	Y
2	Y
3	N
4	Y
5	N

ภาพที่ 2.1 แสดงตัวอย่างการกำหนดคลาสเป้าหมายแบบไบนารีในรูปแบบ single label

multi-label problem	
instance	class
1	A,C
2	B,C
3	B
4	A,B,C
5	A

ภาพที่ 2.2 แสดงตัวอย่างการกำหนดคลาสเป้าหมายแบบหลายค่าในรูปแบบของ single label

multi-label			
instance	**Target A**	**Target B**	**Target C**
1	Y	N	Y
2	N	Y	Y
3	N	Y	N
4	Y	Y	Y
5	Y	N	N

ภาพที่ 2.3 แสดงตัวอย่างการกำหนดคลาสเป้าหมายแบบไบนารี (binary) ในรูปแบบ multi-label

งานวิจัยนี้กำหนดเป้าหมาย จำนวน 3 เป้าหมาย ซึ่งก็คือทุนทั้งหมด 3 ทุน ได้แก่ ทุนได้เปล่า ทุนกู้ยืมเพื่อการศึกษา และทุนขาดแคลนทุนทรัพย์ ซึ่งแต่ละเป้าหมายหรือทุน ไม่มีความสัมพันธ์กันหรือเป็นอิสระต่อกัน (independent) เนื่องจากเงื่อนไขการให้ทุนแก่นักศึกษานั้น นักศึกษา 1 คน สามารถยื่นขอทุนการศึกษาได้หลายทุน และมีโอกาสได้รับทุนการศึกษาหลายทุน หรือไม่ได้รับทุนการศึกษาใดเลย กล่าวคือ เมื่อนักศึกษาได้รับทุนได้เปล่าแล้วยังสามารถยื่นขอทุนการศึกษาประเภทอื่นได้ ซึ่งการได้รับทุนการศึกษาประเภทหนึ่งไม่มีผลต่อการพิจารณาให้ทุนการศึกษาประเภทอื่น นักศึกษาจึงสามารถยื่นขอทุนการศึกษาได้หลายทุน เหตุนี้ผู้วิจัยจึงทำการการกำหนดคลาสเป้าหมายแบบไบนารี (binary) ในรูปแบบ multi-label ดังภาพที่ 2.3 เพื่อคัดเลือกคุณลักษณะสำคัญแบบพลวัตซึ่งทำการแปรผันไปตามเป้าหมายที่ต้องการ เพื่อคัดเลือกคุณลักษณะสำคัญที่ส่งผลต่อการได้รับทุนการศึกษาประเภทนั้น ๆ

1.4 การจัดการความไม่สมดุลของข้อมูล (imbalanced data)

การจัดการความไม่สมดุลของข้อมูล (imbalanced data) เป็นปัญหาที่เกิดขึ้นเมื่อข้อมูลกลุ่มหนึ่งมากกว่าอีกกลุ่มหนึ่งเป็นจำนวนมาก ซึ่งส่งผลทำให้การจำแนกข้อมูลมีความโน้มเอียงไปทางข้อมูลกลุ่มมาก ดังภาพที่ 2.4

class	data
yes (minority class)	40
no (majority class)	960

ภาพที่ 2.4 แสดงความไม่สมดุลของข้อมูล (imbalanced data)

จากภาพที่ 2.4 แสดงให้เห็นว่าข้อมูลการจำแนกประเภทมีความไม่สมดุลของข้อมูล (imbalanced data) เนื่องจากมีข้อมูลกลุ่ม “yes” จำนวน 40 ข้อมูล และข้อมูลกลุ่ม “no” จำนวน 960 ข้อมูล ซึ่งมีความแตกต่างกันเป็นจำนวนมาก เมื่อนำข้อมูลที่มีความไม่สมดุลไปทำการจำแนกประเภทข้อมูลด้วยอัลกอริทึมจำแนกประเภทข้อมูล จะส่งผลทำให้การจำแนกประเภทข้อมูลมีความเอนเอียง (bias) ไปทางกลุ่มข้อมูลที่จำนวนมาก ซึ่งทำให้ข้อมูลที่อยู่ในคลาสกลุ่มน้อยเกิดการจำแนกผิดประเภท ซึ่งอาจส่งผลทำให้ไม่สามารถจำแนกประเภทข้อมูลที่อยู่ในคลาสส่วนน้อยได้

งานวิจัยนี้มีจำนวนข้อมูลการจำแนกประเภทแตกต่างกันจำนวนมาก ซึ่งมีจำนวนการไม่ได้รับทุน มากกว่าการได้รับทุน ดังนั้นผู้วิจัยจึงใช้วิธีการจัดการความไม่สมดุลของข้อมูล (imbalanced data) เพื่อปรับสมดุลของข้อมูลให้มีจำนวนใกล้เคียงกัน เพื่อแก้ปัญหาการจำแนกข้อมูลเอนเอียงไปทางกลุ่มมาก

1.5 การสุ่มตัวอย่างกลุ่มน้อยสังเคราะห์ (Synthetic Minority Oversampling Technique: SMOTE)

วิธีการสุ่มตัวอย่างกลุ่มน้อยสังเคราะห์ (Synthetic Minority Oversampling Technique: SMOTE) (Chawla, 2002) เป็นเทคนิคการปรับเพิ่มจำนวนข้อมูลกลุ่มน้อย ซึ่งเป็นการสุ่มเพิ่มจำนวนข้อมูลกลุ่มน้อย (minority class) ให้มีจำนวนใกล้เคียงกับข้อมูลกลุ่มมาก (majority class) โดย SMOTE เป็นเทคนิคที่ใช้ในการแก้ปัญหาการจำแนกข้อมูลที่ไม่สมดุล เนื่องจากข้อมูลในแต่ละคลาสมีจำนวนแตกต่างกันมาก ทำให้ผลลัพธ์ของการจำแนกข้อมูลอยู่ในกลุ่มมาก ดังนั้นวิธี SMOTE เป็นการเพิ่มจำนวนข้อมูลกลุ่มน้อยให้มีจำนวนเพิ่มขึ้น

โดยการเพิ่มข้อมูลในกลุ่มข้อมูลน้อยนั้นทำให้การกระจายของกลุ่มข้อมูลมีความสมดุลมากขึ้น โดยทำการสุ่มค่าข้อมูลที่อยู่ในกลุ่มข้อมูลน้อยขึ้นมา 1 ค่า หลังจากนั้นพิจารณาค่าข้อมูลใกล้เคียงอีกจำนวน k ค่า แล้วคำนวณค่าระยะทาง (euclidean distance) ระหว่างค่าที่สุ่มกับค่าข้อมูลใกล้เคียงแต่ละค่า เพื่อหาค่าระยะทางที่น้อยที่สุดระหว่างค่าที่สุ่มกับค่าใกล้เคียง จากนั้นจึงสร้างข้อมูล

เทียมระหว่างค่าข้อมูลที่สุ่มกับค่าข้อมูลใกล้เคียงตัวที่ให้ค่าระยะทางที่น้อยที่สุด (He, H. And Garcia, E.A., 2009) ดังสมการที่ 9

$$X_{new} = x_i + (\hat{x}_i - x_i) \times \delta \qquad (9)$$

โดย

X_{new} คือ ข้อมูลใหม่

$\hat{x}_i$ คือ ข้อมูลที่สุ่มในตอนแรก

x_i คือ ข้อมูลที่สุ่มเพิ่มมาอีก

δ คือ ค่าสุ่มตั้งแต่ 0-1

ด้วยงานวิจัยนี้มีจำนวนข้อมูลการจำแนกประเภทแตกต่างกันจำนวนมาก ซึ่งมีจำนวนการไม่ได้รับทุน มากกว่าการได้รับทุน ดังนั้นผู้วิจัยจึงเลือกใช้เทคนิควิธีการสุ่มตัวอย่างกลุ่มน้อยสังเคราะห์ (Synthetic Minority Oversampling Technique: SMOTE) เพื่อปรับเพิ่มข้อมูลให้มีจำนวนใกล้เคียงกัน เพื่อแก้ปัญหาการจำแนกข้อมูลเอนเอียงไปทางกลุ่มมาก ดังภาพที่ 2.5

class	original	SMOTE
yes (minority class)	40	960
no (majority class)	960	960

ภาพที่ 2.5 แสดงการปรับสมดุลของข้อมูลด้วยวิธี SMOTE

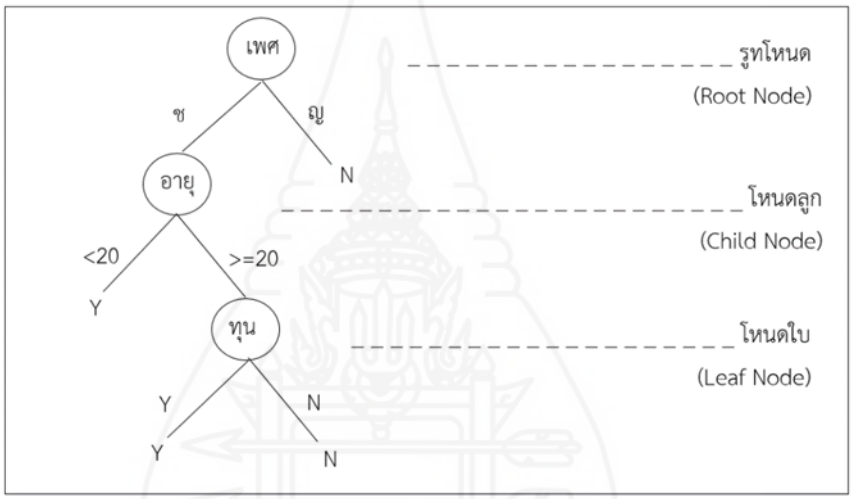
จากภาพที่ 2.5 แสดงให้เห็นว่าข้อมูลการจำแนกประเภทมีความไม่สมดุลของข้อมูล (imbalanced data) เนื่องจากมีข้อมูลกลุ่ม “yes” จำนวน 40 ข้อมูล และข้อมูลกลุ่ม “no” จำนวน 960 ข้อมูล ซึ่งมีความแตกต่างกันเป็นจำนวนมาก เมื่อนำข้อมูลที่มีความไม่สมดุลมาปรับสมดุลด้วยวิธี SMOTE ทำให้จำนวนของข้อมูลกลุ่ม “yes”เพิ่มมากขึ้น เมื่อนำไปทำการจำแนกประเภทข้อมูลด้วยอัลกอริทึมจำแนกประเภทข้อมูล ทำให้สามารถลดความเอนเอียง (bias) ในการจำแนกประเภทข้อมูล

1.6 เทคนิคการจำแนกประเภทข้อมูล (classification)

การจำแนกประเภทข้อมูล (classification) เป็นกระบวนการสร้างโมเดลจัดการข้อมูลให้อยู่ในกลุ่มที่กำหนดมาให้จากกลุ่มตัวอย่างที่เรียกว่า (training data) ที่แต่ละแถวของข้อมูลประกอบด้วยแอตทริบิวต์ (attribute) จำนวนมาก ซึ่งแอตทริบิวต์นี้อาจจะเป็นค่าต่อเนื่อง (continuous) ที่เก็บข้อมูลประเภทตัวเลข (number data) หรือข้อมูลจำแนกประเภท (categorical data) โดยมีแอตทริบิวต์ที่ทำหน้าที่ในการจำแนกประเภทข้อมูล (classification attribute) ซึ่งเป็นตัวกำหนดคลาส (class)

1.7 เทคนิคต้นไม้ตัดสินใจ (decision tree)

เทคนิคต้นไม้ตัดสินใจ (decision tree) เป็นเทคนิคการจำแนกประเภทข้อมูลที่ให้ผลลัพธ์ในลักษณะของโครงสร้างต้นไม้ (Kaewchinporn, 2010) ต้นไม้ตัดสินใจ (ปราโมทย์ ลือนาม, 2561) มีลักษณะเหมือนโครงสร้างต้นไม้ ซึ่งประกอบด้วยโหนดแรกสุดที่เรียกว่า “รูทโหนด” (root node) จากรูทโหนดจะแตกออกเป็นโหนดลูก (child node) โหนดลูกอาจมีลูกของตัวเอง จนถึงโหนดในระดับสุดท้ายจะเรียกว่าโหนดใบ (leaf node) โหนดจะสร้างแอตทริบิวต์ (attribute) เส้นที่เชื่อมโยงแต่ละโหนดเรียกว่ากิ่ง ใช้แสดงเส้นทางที่เป็นไปได้หลังจากตรวจสอบเงื่อนไขของข้อมูล ส่วนโหนดใบใช้สำหรับแสดงกลุ่มหรือคลาส (class) ตามที่กำหนดไว้ โดยเทคนิคต้นไม้ตัดสินใจจะให้ผลลัพธ์เพียง 1 แอตทริบิวต์ผลลัพธ์ (output attribute) เท่านั้น ดังภาพที่ 2.6



ภาพที่ 2.6 แสดงตัวอย่างต้นไม้ตัดสินใจที่ใช้ในการจำแนก

ต้นไม้ตัดสินใจมีการคำนวณค่าของแต่ละคุณลักษณะ ดังค่าต่อไปนี้

ค่า gini index คือ ค่าที่บ่งบอกว่าคุณลักษณะใดควรนำมาใช้เป็นคุณลักษณะในการแบ่งกลุ่มของอัลกอริทึม j48 ดังสมการที่ 10

$$Gini(x_i) = 1 - \sum_{i=1}^{N}[p(t_i)]^2 \qquad (10)$$

ค่า entropy คือ ค่าคาดคะเนของข้อมูลเป็นค่าที่จำแนกโดยใช้ลักษณะประจำของอัลกอริทึม j48 ดังสมการที่ 11

$$Entropy(t_i) = 1 - \sum_{i=1}^{N}[p(t_i)]\log_2 p(t_i) \qquad (11)$$

โดยที่

t_i คือ คุณลักษณะที่นำมาวัดค่า entropy

Py_i คือ สัดส่วนของจำนวนสมาชิกของกลุ่ม i กับจำนวนสมาชิกทั้งหมดของกลุ่มตัวอย่าง

ซึ่งต้นไม้ตัดสินใจมีอัลกอริทึมที่หลากหลาย แต่ละอัลกอริทึมจะให้ผลลัพธ์โครงสร้างต้นไม้ตัดสินใจที่แตกต่างกันไป ในงานวิจัยนี้เลือกใช้เทคนิคต้นไม้ตัดสินใจแบบ j48

1.7.1 อัลกอริทึมของ j48

อัลกอริทึมของ j48 เป็นอัลกอริทึมในการสร้างต้นไม้ตัดสินใจจากกลุ่มข้อมูลฝึกสอนโดยใช้ความถูกต้องของแต่ละคุณลักษณะของข้อมูล เพื่อใช้เป็นการตัดสินใจแบ่งกลุ่มย่อย ๆ โดยพิจารณาจากค่าความแตกต่างในค่า entropy ผลลัพธ์จากการเลือกคุณลักษณะสำหรับแบ่งกลุ่มข้อมูลด้วยค่า normalized Information Gain ที่สูงที่สุดนั้นคือการสร้างการตัดสินใจ (Chawla, 2002)

งานวิจัยนี้เป็นการสร้างโมเดลเพื่อการทำนายการได้รับทุน ซึ่งเป็นการเรียนรู้แบบมีผู้สอน (supervised learning) โดยใช้ชุดข้อมูลฝึกสอน (training data) และข้อมูลผลลัพธ์ (output data) สำหรับการจำแนกประเภทข้อมูล (categorical classification) ดังนั้นผู้วิจัยจึงเลือกใช้เทคนิคต้นไม้ตัดสินใจ (decision tree) ซึ่งมีความเหมาะสมกับข้อมูลการจำแนกประเภท และเป็นการกำหนดคลาสแบบไบนารี (binary) ซึ่งเป็นการจำแนกประเภทข้อมูลออกเป็น 2 คลาส ได้แก่ การได้รับทุนการศึกษา และไม่ได้รับทุนการศึกษาประเภทนั้น ๆ

1.8 การวัดประสิทธิภาพแบบจำลอง (evaluation)

ในกระบวนการขั้นตอนการพัฒนาแบบจำลองและการนำแบบจำลองไปใช้งาน ขั้นตอนการประเมินแบบจำลองเป็นขั้นตอนสำคัญขั้นตอนหนึ่ง เพื่อให้ทราบคุณภาพและประสิทธิภาพของแบบจำลองที่พัฒนาขึ้น โดยสามารถนำแบบจำลองหลาย ๆ แบบ จากเทคนิคต่าง ๆ มาทำการเปรียบเทียบเพื่อเลือกแบบจำลองที่ดีที่สุดในการนำไปใช้งาน โดยมีวิธีการประเมินผลและวัดประสิทธิภาพแบบจำลอง ดังนี้

สุรพงษ์ เอื้อวัฒนามงคล (2557) กล่าวว่า การประเมินและจำลองการจำแนกประเภทข้อมูล (classification) ในการวัดประสิทธิภาพการจำแนกประเภทของแบบจำลองใด ๆ สามารถใช้เครื่องมือหรือมาตรวัดได้หลายอย่าง ดังนี้

1) การทดสอบแบบไขว้ทบ (k – fold cross validation) เป็นวิธีที่ใช้ข้อมูลฝึกสอนบางส่วนโดยแบ่งชุดข้อมูลออกเป็น k กลุ่ม โดยเริ่มต้นแบ่งกลุ่มข้อมูลออกเป็นส่วน ๆ เท่า ๆ กัน จะใช้ข้อมูลในการสร้างตัวแบบ k-1 กลุ่ม และใช้ข้อมูล 1 กลุ่ม เป็นข้อมูลทดสอบ และโดยอาจกำหนดให้ k = n โดยที่ n คือ จำนวนข้อมูลทั้งหมด ซึ่งหมายถึง ในแต่ละครั้งจะใช้ข้อมูล n = 1 ตัวในการสร้างตัวแบบ และใช้ข้อมูลที่เหลือ 1 ตัว เป็นตัวทดสอบ ทำซ้ำจนกระทั่งข้อมูลทุกส่วนถูกนำมาทดสอบ ซึ่งวิธีนี้เรียกอีกอย่างหนึ่งว่า leave – one – out

ในงานวิจัยนี้ได้เลือกใช้ค่า k = 10 หากกำหนดให้ k = 10 กลุ่ม กระบวนการทดสอบจะใช้ข้อมูล 9 กลุ่ม ในการพัฒนาแบบจำลอง และ 1 กลุ่มจะเป็นข้อมูลทดสอบ ทำเช่นนี้ k = 10 ครั้ง รายละเอียดอธิบายดังภาพที่ 2.7

	training set									testing set
รอบที่ 1	2	3	4	5	6	7	8	9	10	1
รอบที่ 2	1	3	4	5	6	7	8	9	10	2
รอบที่ 3	1	2	4	5	6	7	8	9	10	3
รอบที่ 4	1	2	3	5	6	7	8	9	10	4
รอบที่ 5	1	2	3	4	6	7	8	9	10	5
รอบที่ 6	1	2	3	4	5	7	8	9	10	6
รอบที่ 7	1	2	3	4	5	6	8	9	10	7
รอบที่ 8	1	2	3	4	5	6	7	9	10	8
รอบที่ 9	1	2	3	4	5	6	7	8	10	9
รอบที่ 10	1	2	3	4	5	6	7	8	9	10

ภาพที่ 2.7 แสดงการทดสอบแบบไขว้ทบ (k – fold cross validation)

จากภาพที่ 2.7 แสดงการแบ่งชุดข้อมูลออกเป็น 10 กลุ่ม โดยเริ่มต้นแบ่งกลุ่มข้อมูลออกเป็นส่วน ๆ เท่า ๆ กัน จะใช้ข้อมูลกลุ่มที่ 2-10 ในการสร้างตัวแบบ และใช้ข้อมูลกลุ่ม 1 เป็นข้อมูลทดสอบ ทำซ้ำจนกระทั่งครบ

2) Confusion matrix การคำนวณประสิทธิภาพของแบบจำลอง สามารถคำนวณได้ตามตาราง confusion matrix ซึ่งเป็นตารางสรุปจำนวนข้อมูลที่มีการจำแนกได้ถูกต้องและไม่ถูกต้อง (Witten, 2011) ตัวอย่างดังตารางที่ 2.2

ตารางที่ 2.2 ตาราง confusion matrix

		Predicted	
		Positive	**Negative**
Class predict	**Positive**	**TP**	**FN**
	Negative	**FP**	**TN**

3) ค่าความถูกต้อง (accuracy) เป็นเกณฑ์วัดความถูกต้องหรือความแม่นยำ ในการจำแนกประเภทเพื่อบ่งบอกระดับความถูกต้องในการจำแนกประเภทข้อมูล แบบจำลองที่ได้จากการประมวลผล ได้แก่ สัดส่วนระหว่างจำนวนข้อมูลทั้งหมดที่จำแนกประเภทความถูกต้องทั้งประเภท positive และ negative กับจำนวนข้อมูลทั้งหมดที่มีถูกจำแนกประเภทดังสมการที่ (12)

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (12)$$

4) ค่าความแม่นยำ (precision) เป็นการวัดความสามารถของระบบในการขจัดคำตอบหรือ ข้อมูลที่ไม่เกี่ยวข้องออกไป ถ้าระบบสามารถขจัดคำตอบหรือข้อมูลที่ไม่เกี่ยวข้องออกไปได้มาก แสดงถึงความแม่นยำของระบบสูง ดังสมการที่ (13)

$$Precision = \frac{TP}{TP+FP} \quad (13)$$

5) ค่าเรียกคืน (recall) เป็นการวัดค่าตัวแบบจำลอง โดยพิจารณาแยกทีละผลลัพธ์ที่สนใจ ซึ่งเป็นการบอกว่าแบบจำลองจำแนกประเภทมีค่าความถูกต้องในการทำนายค่าที่ถูกต้องจากค่าที่ถูกต้องจริงทั้งหมดมากน้อยเพียงใด ดังสมการที่ (14)

$$Recall = \frac{TP}{TP+FN} \quad (14)$$

6) ค่าประสิทธิภาพโดยรวม (f-measure) เป็นการวัดค่า precision และค่า recall พร้อมกันของตัวแบบจำลองจำแนกประเภท โดยพิจารณาแยกทีละผลลัพธ์ที่สนใจและไม่สนใจ ซึ่งเป็นค่าที่ใช้สำหรับบอกความมีประสิทธิภาพของแบบจำลองการจำแนกประเภท ว่ามีความเหมาะสมในการนำไปใช้กับข้อมูลอื่น ๆ ที่เป็นข้อมูลใหม่ได้ดีเพียงใด ดังสมการที่ (15)

$$F - measure = \frac{2x\ Precision\ x\ Recall}{Precision + Recall} \quad (15)$$

โดย TP (True Positive) คือ จำนวนข้อมูลบอกว่าจริง และพยากรณ์ว่าจริง
TN (True Negative) คือ จำนวนข้อมูลบอกว่าไม่จริง และพยากรณ์ว่าไม่จริง
FP (False Positive) คือ จำนวนข้อมูลบอกว่าไม่จริง แต่พยากรณ์ว่าจริง
FN (False Negative) คือ จำนวนข้อมูลบอกว่าจริง แต่พยากรณ์ว่าไม่จริง

2. งานวิจัยที่เกี่ยวข้อง

2.1 งานวิจัยในประเทศ

นิภาพร ชนะมาร และพรรณี สิทธิเดช (2557) ได้วิจัย “การวิเคราะห์ปัจจัยการเรียนรู้ด้วยการคัดเลือกคุณสมบัติและการพยากรณ์” เพื่อประยุกต์ใช้เทคนิคเหมืองข้อมูลทำนายผลสัมฤทธิ์ทางการเรียนของนิสิตระดับปริญญาตรี สาขาวิชาวิทยาการคอมพิวเตอร์ มหาวิทยาลัยนเรศวร จำนวน 180 ระเบียน ประกอบด้วยคุณสมบัติ 23 คุณลักษณะ แบ่งเป็นตัวแปรอิสระ 22 คุณลักษณะ ได้แก่ ข้อมูลภูมิหลังและข้อมูลผลการเรียนรายวิชาในชั้นปีที่ 1 และ 2 ตัวแปรตามหรือตัวแปรทำนาย คือ เกรดเฉลี่ยเมื่อสำเร็จการศึกษา จากนั้นวิเคราะห์ปัจจัยการเรียนรู้ด้วยการคัดเลือกคุณลักษณะที่สำคัญ 3 วิธี ได้แก่ 1) วิธี correlation-based 2) consistency-based และ 3) วิธี Gain Ratio จากนั้นนำมาพัฒนาแบบจำลองการทำนาย ด้วยเทคนิค neural network แบบ back-propagation และเทคนิค support vector machine ร่วมกับการวัดประสิทธิภาพด้วยวิธี 10-fold cross validation และวัดค่ารากที่สองของความคลาดเคลื่อน (RMSE) พบว่า ข้อมูลภูมิหลังไม่ใช่ข้อมูลสำคัญในการทำนายผลสัมฤทธิ์ทางการเรียน และผลการเรียนรายวิชา จำนวน 10 คุณลักษณะนั้นเป็นตัวแปรสำคัญ สำหรับแบบจำลองทำนายด้วยเทคนิค neural network แบบ back-propagation และเทคนิค support vector machine พบว่า ผลการพยากรณ์ของ bagging ร่วมกับเทคนิค neural network แบบ back-propagation มีค่ารากที่สองของความคลาดเคลื่อน (RMSE) อยู่ระดับต่ำที่สุดที่ 0.1051 มีประสิทธิภาพดีที่สุดจึงนำไปใช้ในการพยากรณ์

ภรัณยา ปาลวิสุทธิ์ (2559) ได้วิจัย “การเพิ่มประสิทธิภาพเทคนิคต้นไม้ตัดสินใจบนชุดข้อมูลที่ไม่สมดุล โดยวิธีการสุ่มเพิ่มตัวอย่างกลุ่มน้อยสำหรับข้อมูลการเป็นโรคติดอินเทอร์เน็ต” เพื่อที่จะพัฒนาตัวแบบสำหรับพยากรณ์การเป็นโรคติดอินเทอร์เน็ตในเยาวชน ซึ่งผลการวิเคราะห์ข้อมูลพบว่า ข้อมูลมีความไม่สมดุลของกลุ่มข้อมูล โดยมีจำนวนกลุ่มหนึ่งมากกว่าอีกกลุ่มหนึ่งเป็นจำนวนมาก จึงได้ทำการแก้ไขปัญหาโดยการปรับสมดุลของข้อมูลด้วยวิธีเทคนิคการสุ่มเพิ่มตัวอย่างกลุ่มน้อย (Synthetic Minority Over-Sampling Technique: SMOTE) แล้วพัฒนาตัวแบบด้วยเทคนิคต้นไม้ตัดสินใจ j48, id3, lmt, cart และ random forest โดยใช้ 10-fold cross validation ในการแบ่งข้อมูลออกเป็นชุดข้อมูลสอนและชุดข้อมูลทดสอบ และใช้ค่าความแม่นยำ (accuracy) ค่าความไว (sensitive) และค่าความจำเพาะ (specificity) ในการวัดประสิทธิภาพการพยากรณ์ของตัวแบบ ผลการทดลองได้ประสิทธิภาพการพยากรณ์ของตัวแบบ พบว่า เทคนิค random forest สามารถพยากรณ์ได้ดี ซึ่งมีค่าความแม่นยำร้อยละ 87.15 ค่าความไวร้อยละ 85.89 และค่าความจำเพาะร้อยละ 87.53

รัชพล กลัดชื่น และจรัญ แสนราช (2561) ได้วิจัย “การเปรียบเทียบประสิทธิภาพอัลกอริทึมและการคัดเลือกคุณลักษณะที่เหมาะสมเพื่อทำนายผลสัมฤทธิ์ทางการเรียนของนักศึกษาระดับอาชีวศึกษา” เพื่อเปรียบเทียบประสิทธิภาพของอัลกอริทึมในการทำนายและคุณลักษณะที่มีต่อผลสัมฤทธิ์ทางการเรียนของนักศึกษาระดับอาชีวศึกษา โดยทำการศึกษาข้อมูลนักศึกษาระดับประกาศนียบัตรวิชาชีพ จำนวน 5,100 ระเบียน ตั้งแต่ปีการศึกษา 2550 – 2559 จำนวน 9 สาขาวิชา และจำนวน 27 คุณลักษณะ โดยใช้เทคนิคการจำแนกข้อมูล 3 เทคนิค ได้แก่ เทคนิค decision tree โดยใช้อัลกอริทึม j48 graft เทคนิค naïve bayes และเทคนิค rule induction ทำการเปรียบเทียบประสิทธิภาพตัวแบบการทำนายระหว่างการใช้คุณลักษณะทั้งหมดกับการเลือกคุณลักษณะแบบ Forward Selection ทดสอบประสิทธิภาพตัวแบบทำนายด้วยวิธีการ 10-fold cross validation โดยใช้โปรแกรม rapid miner studio 8 จากนั้นนำผลการทดสอบประสิทธิภาพที่มีค่าความถูกต้องที่สูงที่สุด 2 ค่า มาทำการเปรียบเทียบด้วยวิธี t-test ผลการศึกษาพบว่า การใช้เทคนิค decision tree : j48 graft ด้วยการคัดเลือกแบบ Forward Selection และการคัดเลือกคุณลักษณะทั้งหมด มีค่าความถูกต้องร้อยละ 83.08 และร้อยละ 81.71 ตามลำดับ และทดสอบด้วยวิธี t-test พบว่าการทดสอบทั้งสองแบบมีความแตกต่างกันอย่างมีนัยสำคัญทางสถิติที่ระดับ 0.05

พุทธิพร ธนธรรมเมธี และเยาวเรศ ศิริสถิตกุล (2561) ได้วิจัยเรื่อง “เทคนิคการจำแนกข้อมูลที่พัฒนาสำหรับข้อมูลที่ไม่สมดุลของภาวะข้อเข่าเสื่อมในผู้สูงอายุ” เพื่อพัฒนาตัวแบบการพยากรณ์ภาวะข้อเข่าเสื่อมในผู้สูงอายุ จากข้อมูลแบบบันทึกการประเมินข้อเข่าเสื่อม โรงพยาบาลส่งเสริมสุขภาพตำบลบ้านหาร อำเภอท่าศาลา จังหวัดนครศรีธรรมราช จำนวน 370 เรคอร์ด และมีข้อมูล 4 คลาส ได้แก่คลาส 0 ยังไม่พบอาการผิดปกติ 200 เรคอร์ด คลาส 1 เริ่มมีอาการข้อเข่าเสื่อม 115 เรคอร์ดคลาส 2 มีอาการโรคข้อเข่าเสื่อมระดับปานกลาง 39 เรคอร์ด และคลาส 3 เป็นโรคข้อเข่าเสื่อมระดับรุนแรง 16 เรคอร์ด สำหรับการวินิจฉัยทางการแพทย์ ข้อมูลกลุ่มน้อย คือ ข้อมูลที่สนใจและการจำแนกผิดพลาดเกิดขึ้นได้สูงกว่าข้อมูลกลุ่มมาก ซึ่งข้อมูลชุดนี้มีจำนวนรวมของคลาส 0 และคลาส 1 สูงกว่าคลาส 2 และคลาส 3 เป็นจำนวนมาก จึงเกิดความไม่สมดุลของข้อมูล ส่งผลให้การจำแนกข้อมูลผิดพลาดได้การปรับความไม่สมดุลของข้อมูลคลาส 2 และคลาส 3 ทำได้ด้วยเทคนิคการปรับเพิ่มข้อมูลด้วยวิธีสุ่ม โดยใช้วิธี ADASYN และ SMOTE และใช้งานวิธีการตรวจสอบไขว้แบบ 10 กลุ่ม ในการแบ่งเป็นชุดข้อมูลสอนและชุดข้อมูลทดสอบ จากนั้นจำแนกข้อมูลด้วย multi-class imbalanced data classification ด้วยวิธี one-vs-one และ one-vs-all และเทคนิค gentleboost ผลการทดสอบประสิทธิภาพพบว่าวิธี ADASYN และ one-vs-one ให้ค่าความถูกต้อง 97.31% และทดสอบตัวแบบกับชุดข้อมูลจริงที่ไม่สมดุลจาก โรงพยาบาลส่งเสริมสุขภาพตำบลบ้านหัวคูอำเภอท่าศาลาจังหวัดนครศรีธรรมราชจำนวน 232 เรคอร์ด และมีข้อมูล 4 คลาส ได้แก่ คลาส 0 ยังไม่พบอาการผิดปกติ141เรคอร์ด คลาส 1 เริ่มมีอาการข้อเข่าเสื่อม 63 เรคอร์ด คลาส 2 มีอาการโรคข้อเข่า

เสื่อมระดับปานกลาง 16 เรคอร์ดและคลาส 3 เป็นโรคข้อเข่าเสื่อมระดับรุนแรง 12 เรคอร์ด พบว่าจำแนกถูกต้อง 85.78 % และจำแนกคลาส 2 และคลาส 3 ได้ถูกต้องเพิ่มขึ้น โดยเฉพาะคลาส 3 เพิ่มขึ้นจาก 0 เป็น 75 % ซึ่งตัวแบบนี้สามารถนำมาใช้ในแผนส่งเสริมสุขภาพเพื่อวินิจฉัยและบำบัดผู้สูงอายุ

กาญจน์ ณ ศรีธะ และคณะ (2561) ได้วิจัย “การเปรียบเทียบเทคนิคการสุ่มตัวอย่างเพื่อการจำแนกข้อมูลที่ไม่สมดุล” เพื่อเปรียบเทียบเทคนิคการสุ่มตัวอย่างเพื่อการจำแนกข้อมูลที่ไม่สมดุล โดยได้ทำการทดลองกับข้อมูลจำนวน 3 ชุดข้อมูล ใช้เทคนิคการสุ่มเพิ่มตัวอย่างส่วนน้อย เทคนิคการสุ่มลดตัวอย่างกลุ่มมาก และเทคนิคการสุ่มตัวอย่างซ้ำสำหรับการปรับปรุงข้อมูลที่ไม่สมดุล ใช้เทคนิคต้นไม้ตัดสินใจ คาร์ท เรนดอมฟรอเรส ซัพพอรท์เวกเตอร์แมชชีน และโครงข่ายประสาทเทียม ร่วมกับเทคนิครวมกลุ่มเอดาบูทและถุงจำแนก เพื่อสร้างแบบจำลองการจำแนกข้อมูล ใช้วิธี 10-fold cross validation เพื่อวัดประสิทธิภาพของแบบจำลอง วัดค่าประสิทธิภาพด้วยค่าความถูกต้อง ค่าความระลึก และค่าเอฟ ผลการวิจัยพบว่า เทคนิคการสุ่มตัวอย่างซ้ำสามารถปรับปรุงข้อมูลที่ไม่สมดุลได้ดีกว่าเทคนิคการสุ่มเพิ่มตัวอย่างกลุ่มน้อย นอกจากนี้ยังพบว่า แบบจำลองเรนดอมฟรอเรส แบบจำลองเอดาบูทร่วมกับเรนดอมฟรอเรส และแบบจำลองถุงจำแนกร่วมกับเรนดอมฟรอเรสมีค่าประสิทธิภาพการจำแนกที่ดีกับข้อมูลในงานวิจัย

อัจจิมา มณฑาพันธุ์ (2562) ได้วิจัย “การเปรียบเทียบการคัดเลือกคุณลักษณะที่สำคัญในการปรับปรุงการพยากรณ์มะเร็งเต้านม” เพื่อศึกษาการคัดเลือกคุณลักษณะที่สำคัญที่ใช้ในการวิเคราะห์ข้อมูลเพื่อปรับปรุงการพยากรณ์การเป็นมะเร็งเต้านม โดยทำการศึกษาเปรียบเทียบเทคนิคการคัดเลือกคุณลักษณะที่สำคัญจำนวน 7 เทคนิค ได้แก่ เทคนิค Correlation base Feature Selection เทคนิค Information Gain เทคนิค Gain Ratio เทคนิค Chi Square เทคนิค Forward Selection เทคนิค Backward Elimination และเทคนิค Evolutionary Selection และใช้ข้อมูลในการวิเคราะห์จาก UCI machine learning repository เป็นคนที่ยังมีชีวิตอยู่จำนวน 569 คน และทดสอบประสิทธิภาพตัวแบบทำนายด้วยวิธีการ 10-fold cross validation จากการวัดค่าประสิทธิภาพตัวแบบด้วยค่าความถูกต้อง (accuracy) โดยใช้โปรแกรม rapid miner studio พบว่าหากใช้ข้อมูลทั้งหมดโดยไม่ลดมิติข้อมูลจะได้ค่าความถูกต้องการพยากรณ์การเป็นมะเร็งเต้านมร้อยละ 91.39 ขณะที่ใช้เทคนิค Evolutionary Selection สามารถคัดเลือกคุณลักษณะสำคัญเพื่อใช้ในการวิเคราะห์ข้อมูลเพียง 16 คุณลักษณะ และให้ค่าความถูกต้องการพยากรณ์การเป็นมะเร็งเต้านมได้ดีที่สุดร้อยละ 95.26 ซึ่งพบว่า การคัดเลือกคุณลักษณะที่สำคัญด้วยเทคนิค Evolutionary Selection สามารถลดมิติของข้อมูลจาก 30 คุณลักษณะ ลงมาเหลือเพียง 16 คุณลักษณะ และสามารถเพิ่มความแม่นยำในการวัดค่าความถูกต้องจากร้อยละ 91.39 มาเป็นร้อยละ 95.25

วิชญ์วิสิฐ เกษรสิทธิ์ และคณะ (2563) ได้วิจัยเรื่อง “การลดจำนวนกลุ่มในการจำแนกแบบหลายกลุ่มเป็นสองกลุ่มสำหรับการจำแนกการกลับมารักษาซ้ำในโรงพยาบาลของผู้ป่วยโรคเบาหวาน” การวิจัยครั้งนี้เป็นการวิจัยเพื่อเปรียบเทียบประสิทธิภาพการจำแนกประเภทของการกลับมารักษาซ้ำในโรงพยาบาลของผู้ป่วยโรคเบาหวานแบบหลายกลุ่ม (multiclass) หรือ อเนกนาม (multinomial) และแบบสองกลุ่มหรือทวิภาค (binary) จำนวน 2 กรณี โดยใช้เทคนิคการวิเคราะห์การถดถอยลอจิสติก และต้นไม้การตัดสินใจ ข้อมูลที่ใช้ในการวิจัยเป็นข้อมูลประวัติการรักษาพยาบาลของผู้ป่วยโรคเบาหวานจาก Clinical care at 130 Us Hospitals and Integrated Delivery Networks ตัวแปรเป้าหมายในการจำแนกประกอบด้วยประเภทการนัดหมายให้กลับมารักษาซ้ำในโรงพยาบาลของผู้ป่วยโรคเบาหวาน จำนวน 3 กลุ่ม คือ ไม่กลับมารักษาซ้ำหรือไม่มีภาวะโรคกลับมารักษาซ้ำภายใน 30 วัน และกลับมารักษาซ้ำมากกว่า 30 วัน ผลการวิจัยพบว่าประสิทธิภาพของการจำแนกประเภทโดยใช้เทคนิคต้นไม้การตัดสินใจแบบทวิภาค จำนวน 2 กรณีมีประสิทธิภาพสูงสุด

วรการ ใจดี และนพณัฐ วรรณภีร์ (2563) ได้วิจัยเรื่อง “การศึกษาปัจจัยที่ส่งผลต่อการสำเร็จการศึกษาตามแผนของนักศึกษาระดับปริญญาตรี โดยใช้เทคนิคการคัดเลือกคุณลักษณะบนชุดข้อมูลที่ไม่สมดุล” เพื่อศึกษาปัจจัยที่ส่งผลต่อการสำเร็จการศึกษาตามแผนของนักศึกษา และเพื่อศึกษาแนวทางและข้อปรับปรุงในการแก้ไขปัญหาการเรียนการสอนในแต่ละรายวิชาที่ส่งผลต่อการสำเร็จการศึกษาตามแผนของนักศึกษา จากการรวบรวมข้อมูลพบว่า ชุดข้อมูลมีความไม่สมดุลของข้อมูล (imbalanced datasets) โดยมีจำนวนกลุ่มนักศึกษาที่สำเร็จการศึกษาตามแผนมากกว่ากลุ่มนักศึกษาที่ไม่สำเร็จกรศึกษาตามแผน จึงใช้ทำการแก้ปัญหาโดยใช้วิธีการปรับสมดุลให้กับชุดข้อมูลโดยใช้เทคนิคการสุ่มตัวอย่างข้อมูลกลุ่มน้อย (Synthetic Minarity Over-Sampling Technique : SMOTE) ก่อนที่จะนำข้อมูลเข้าสู่กระบวนการวิเคราะห์ปัจจัยโดยใช้เทคนิคการคัดเลือกคุณลักษณะ (feature selection) จำนวน 3 เทคนิค คือ เทคนิค Chi-Square Feature Selection เทคนิค Information Gain Feature Selection และเทคนิค Correlation based Feature Selection

2.2 งานวิจัยต่างประเทศ

Khan et al. (2011) ได้ศึกษาวิจัยเพื่อแก้ปัญหาความไม่สมดุลของชุดข้อมูลสำหรับข้อมูลทางด้านชีววิทยา โดยได้นำหลัดการทำงานแบบดั้งเดิมของ machine learning ที่เรียกว่า support vector machine มาทำการปรับปรุงสำหรับวิเคราะห์ชุดข้อมูลทางด้าน eukaryyotic genomes จากผลการวิจัยพบว่า จากชุดข้อมูลที่ไม่สมดุลที่มีอัตราส่วน 1:4500 ซึ่งวิธีการทาง mchine learning แบบดั้งเดิมไม่สามารถดำเนินการได้ดีนั้น เมื่อใช้วิธีการที่นำเสนอ โดยใช้หลักการทำ under sampling กลุ่มข้อมูลหลัก (majority class) และใช้วิธีการแบบ heuristics เพื่อลด

สัดส่วนความไม่สมดุลของข้อมูล หลังจากนั้นใช้วิธีการ SMOTE เพื่อสร้างคุณลักษณะที่ต้องการ (desired feature) และทำการเลือกคุณลักษณะที่ดีที่สุดด้วย feature selection หลาย ๆ ตัว กับชุดข้อมูลตัวอย่างของสายพันธ์ DNA ซึ่งมีทั้งหมด 11,120 ลักษณะ ให้เหลือเพียง 15 ลักษณะ และใช้หลักการความคล้ายคลึงกัน (similarity) ในการสร้างชุดทดสอบ ผลหารทดลองพบว่า วิธีการที่นำเสนอให้ค่า f-measure อยู่ที่ระดับ 0.44 สัดส่วนของค่า recall และค่า precision เท่ากับ 15%100% 29%85% และ 85%4%

Osiris Villacampa (2015) ได้วิจัยเรื่อง “feature selection and classification methods for decision making: a comparative analysis” เพื่อศึกษาการประมวลผลเพื่อลดมิติข้อมูลด้วยการคัดเลือกคุณลักษณะหรือการเลือกคุณลักษณะซึ่งจะช่วยให้มีความถูกต้องและมีประสิทธิภาพ โดยใช้ข้อมูลประวัติการบริการและการขายรถจากตัวแทนจำหน่าย จำนวน 15,415 ระเบียน จำนวน 40 คุณลักษณะ และประวัติการเงินของลูกค้าจากธนาคาร จำนวน 10,578 ระเบียน จำนวน 17 คุณลักษณะ เพื่อหารูปแบบจำแนกประเภทลูกค้าที่มีฐานะอยู่ในกลุ่มผู้ซื้อรถใหม่ โดยนำเทคนิคการคัดเลือกคุณลักษณะแบบ filter, wrapper และ hybrid ได้แก่ วิธี Information Gain วิธี correlation based feature selection วิธี relief-f และวิธี wrappers นำมาใช้เพื่อลดจำนวนคุณลักษณะในชุดข้อมูล จากนั้นทำกรจำแนกประเภทและพัฒนาแบบจำลองด้วยเทคนิค decision tree เทคนิค k-nearest neighbor และเทคนิค support vector machines ทดสอบประสิทธิภาพแบบจำลองด้วยวิธี 5-fold cross validation และวิเคราะห์แปรียบเทียบการคัดเลือกคุณลักษณะและความแตกต่างด้วยค่า accuracy, ค่า area inder receiver operating characteristic cure (auc), ค่า f-measure, ค่า TP rate และค่า FP rate พบว่า ค่าความถูกต้องของเทคนิค decision tree ร่วมกับวิธีการคัดเลือกคุณลักษณะสำคัญแบบ wrapper ที่ค่าความเชื่อมั่น 0.1 มีค่าความถูกต้องแม่นยำร้อยละ 86.40 ค่า auc สูงสุดที่ร้อยละ 90.40 ส่วนเทคนิค k-nearest neighbor และเทคนิค support vector machines มีค่าความถูกต้องเพียงร้อยละ 84.90 และร้อยละ 79.80 ตามลำดับ

จากการศึกษางานวิจัยที่เกี่ยวข้องต่าง ๆ พบว่า งานวิจัยส่วนใหญ่นำข้อมูลที่มีอยู่แล้วมาสร้างโมเดลในการพยากรณ์ผลลัพธ์ โดยส่วนใหญ่ให้ความสำคัญกับการคัดเลือกคุณลักษณะสำคัญ ซึ่งเป็นการคัดเลือกเซตคุณลักษณะเฉพาะใหม่จากเซตคุณลักษณะเดิม โดยเป็นชุดข้อมูลย่อย (data subset) ของชุดข้อมูลเดิม เนื่องจากบางคุณลักษณะเดิมอาจไม่มีความสำคัญหรือมีความสำคัญน้อยถูกคัดเลือกออกไป และด้วยงานวิจัยนี้มีจำนวนข้อมูลการจำแนกประเภทแตกต่างกันจำนวนมาก จากปัจจัยที่ท้าทายคือ จำนวนการไม่ได้รับทุน มากกว่าการได้รับทุน ผู้วิจัยจึงประยุกต์ใช้วิธีการเตรียมข้อมูลด้วยเทคนิคการสุ่มตัวอย่างเพื่อแก้ปัญหาข้อมูลที่ไม่สมดุล เพื่อเพิ่มประสิทธิภาพในการจำแนกข้อมูล โดยเลือกใช้เทคนิคการสุ่มเพิ่มตัวอย่างส่วนน้อยสังเคราะห์ เพื่อปรับเพิ่มข้อมูลให้มี

จำนวนใกล้เคียงกัน เพื่อแก้ปัญหาการจำแนกข้อมูลเอนเอียงไปทางกลุ่มมาก และงานวิจัยนี้กำหนดเป้าหมาย จำนวน 3 เป้าหมาย ซึ่งก็คือทุนทั้งหมด 3 ทุน ได้แก่ ทุนได้เปล่า, ทุนกู้ยืมเพื่อการศึกษา และทุนขาดแคลนทุนทรัพย์ ซึ่งแต่ละเป้าหมายหรือทุน ไม่มีความสัมพันธ์กันหรือเป็นอิสระต่อกัน (independent) จึงทำการคัดเลือกคุณลักษณะสำคัญแบบพลวัต ซึ่งแปรผันไปตามเป้าหมายที่ต้องการ โดยเลือกใช้วิธีการคัดเลือกคุณลักษณะสำคัญที่หลากหลาย จำนวน 7 เทคนิค ได้แก่ เทคนิค Correlation base Feature Selection เทคนิค Information Gain เทคนิค Gain Ratio เทคนิค Chi Square เทคนิค Forward Selection เทคนิค Backward Elimination และเทคนิค Evolutionary Selection เพื่อให้ได้เทคนิคที่มีความเหมาะสมที่สุดสำหรับการจำแนกประเภทต้นไม้ตัดสินใจบนพื้นฐานเงื่อนไขหลายเป้าหมาย ซึ่งงานวิจัยนี้มีความแตกต่างจากงานวิจัยที่ได้ศึกษามาก่อนหน้าคือ ผู้วิจัยทำการเลือกคุณลักษณะสำคัญแบบพลวัตสำหรับความเหมาะสมที่สุดของการจำแนกต้นไม้ตัดสินใจบนพื้นฐานเงื่อนไขหลายเป้าหมาย ซึ่งมีความแตกต่างจากงานวิจัยที่ผ่านมาเป็นการทดลองสร้างโมเดลพยากรณ์ผลลัพธ์เพียงเป้าหมายเดียว เนื่องจากงานวิจัยนี้ทำการบูรณาการการเลือกคุณลักษณะสำคัญแบบพลวัตกับการจำแนกต้นไม้ตัดสินใจบนพื้นฐานเงื่อนไขหลายเป้าหมาย ซึ่งเป็นการคัดเลือกคุณลักษณะสำคัญที่แปรผันไปตามเป้าหมาย และสร้างแบบจำลองการจำแนกผลลัพธ์หลายเป้าหมาย และงานวิจัยนี้เลือกใช้การวัดประสิทธิภาพด้วยมาตรวัด ได้แก่ ค่าความถูกต้อง (accuracy) ค่าความแม่นยำ (precision) ค่าเรียกคืน (recall) และค่าประสิทธิภาพโดยรวม (f-measure) ผู้วิจัยได้ทำการเปรียบเทียบงานวิจัยที่เกี่ยวข้องดังตารางที่ 2.3 – 2.4

ตารางที่ 2.3 ตารางสรุปเปรียบเทียบงานวิจัยที่เกี่ยวข้อง

ขั้นตอนการดำเนินงาน	งานวิจัยที่เกี่ยวข้อง										
	ก	ข	ค	ง	จ	ฉ	ช	ซ	ฌ	ญ	ฎ
1.การจัดการข้อมูล											
SMOTE		√		√	√			√	√		√
ADASYN				√							
2. การคัดเลือกคุณลักษณะสำคัญ (feature selection)											
Correlation base Feature Selection	√					√		√		√	√
Information Gain						√		√		√	√
Gain Ratio	√					√					√
Chi Square						√		√			√
Forward Selection			√			√					√
Backward Elimination						√					√
Evolutionary Selection						√					√
Consistency Base Feature Selection	√										
Relief-F										√	
3. การสร้างโมเดลจำแนกประเภทข้อมูล (classification model)											
Logistic Regression Analysis							√				
Decision Tree		√	√	√	√		√			√	√
Support Vector Machine	√				√	√			√	√	
Naïve Bayes			√								
K-Nearest Neighbor										√	
Neural Network	√				√						
Rule Induction			√								
4. การวัดประสิทธิภาพโมเดล (evaluation)											
k-fold cross validation	√	√	√	√	√	√	√			√	√
Accuracy		√	√	√	√	√	√			√	√
Precision									√		√
RMSE	√										

ตารางที่ 2.3 ตารางสรุปเปรียบเทียบงานวิจัยที่เกี่ยวข้อง (ต่อ)

ขั้นตอนการดำเนินงาน	งานวิจัยที่เกี่ยวข้อง										
	ก	ข	ค	ง	จ	ฉ	ช	ซ	ฌ	ญ	ฎ
Recall									√		√
F-measure									√	√	√
Sensitive		√									
Specification		√									
AUC										√	
TP Rate & FP Rate										√	

หมายเหตุ ก หมายถึง งานวิจัยของนิภาพร ชนะมาร และพรรณี สิทธิเดช (2557)

ข หมายถึง งานวิจัยของภรัณยา ปาลวิสุทธิ์ (2559)

ค หมายถึง งานวิจัยของรัชพล กลัดชื่น และจรัญ แสนราช (2561)

ง หมายถึง งานวิจัยของพุทธิพร ธนธรรมเมธี และเยาวเรศ ศิริสถิตกุล (2561)

จ หมายถึง งานวิจัยของกาญจน์ ณ ศรีธะ และคณะ (2561)

ฉ หมายถึง งานวิจัยของอัจจิมา มณฑาพันธุ์ (2562)

ช หมายถึง งานวิจัยของวิชญ์วิสิฐ เกษรสิทธิ์ และคณะ (2563)

ซ หมายถึง งานวิจัยของวรการ ใจดี และนพณัฐ วรรณภีร์ (2563)

ฌ หมายถึง งานวิจัยของ Khan et al. (2011)

ญ หมายถึง งานวิจัยของ Osiris Villacampa (2015)

ฎ หมายถึง งานวิจัยเรื่อง “การเลือกคุณลักษณะสำคัญแบบพลวัตสำหรับความเหมาะสมที่สุดของการจำแนกต้นไม้ตัดสินใจบนพื้นฐานเงื่อนไขหลายเป้าหมาย” (งานวิจัยของวิทยานิพนธ์นี้)

ตารางที่ 2.4 ตารางเปรียบเทียบรายละเอียดงานวิจัยที่เกี่ยวข้อง

ลำดับ	ชื่องานวิจัย	ผู้วิจัย	ปี	การจัดการข้อมูล	การคัดเลือกคุณลักษณะ	อัลกอริทึมจำแนก	การประเมินประสิทธิภาพ	ผลการวิจัย
ก	การวิเคราะห์ปัจจัยการเรียนรู้ด้วยการคัดเลือกคุณสมบัติและการพยากรณ์	นิภาพร ชนะมาร พรรณี สิทธิเดช	2557		- correlation-based feature selection - consistency-based - Gain Ratio	- neural network แบบ back-propagation - support vector machines	- 10 fold cross-validation - rmse	วิธี bagging ร่วมกับ neural network แบบ back-propagation
ข	การเพิ่มประสิทธิภาพเทคนิคต้นไม้ตัดสินใจบนชุดข้อมูลที่ไม่สมดุลโดยวิธีการสุ่มเพิ่มตัวอย่างกลุ่มน้อยสำหรับข้อมูลการเป็นโรคติดอินเทอร์เน็ต	ภรันยา ปาล วิสุทธิ์	2559	SMOTE	ไม่มีการคัดเลือกคุณลักษณะ	- j48 - id3 - lmt - cart - random forest	- 10-fold cross-validation - accuracy - sensitive - specificity	เทคนิคการสุ่มเพิ่มตัวอย่างกลุ่มน้อย (Synthetic Minority Over-Sampling Technique: SMOTE) และเทคนิค random forest

ตารางที่ 2.4 ตารางเปรียบเทียบรายละเอียดงานวิจัยที่เกี่ยวข้อง (ต่อ)

ลำดับ	ชื่องานวิจัย	ผู้วิจัย	ปี	การจัดการข้อมูล	การคัดเลือกคุณลักษณะ	อัลกอริทึมจำแนก	การประเมินประสิทธิภาพ	ผลการวิจัย
ค	การเปรียบเทียบประสิทธิภาพอัลกอริทึมและการคัดเลือกคุณลักษณะที่เหมาะสมเพื่อทำนายผลสัมฤทธิ์ทางการเรียนของนักศึกษาระดับอาชีวศึกษา	รัชพล กลัดชื่น และ จรัญ แสนราช	2561		- Forward Selection	- decision tree อัลกอริทึม j48 graft - naïve bayes - rule induction	- 10-fold cross validation - t-test - accuracy	วิธี Forward Selection ร่วมกับเทคนิค decision tree ด้วยอัลกอริทึม j48 graft
ง	เทคนิคการจำแนกข้อมูลที่พัฒนาสำหรับข้อมูลที่ไม่สมดุลของภาวะข้อเข่าเสื่อมในผู้สูงอายุ	พุทธิพร ธนธรรมเมธี และเยาวเรศ ศิริสถิตกุล	2561	SMOTE ADASYN	ไม่มีการคัดเลือกคุณลักษณะ	- decision tree	- 10-fold cross-validation - accuracy	เทคนิค decision tree

ตารางที่ 2.4 ตารางเปรียบเทียบรายละเอียดงานวิจัยที่เกี่ยวข้อง (ต่อ)

ลำดับ	ชื่องานวิจัย	ผู้วิจัย	ปี	การจัดการข้อมูล	การคัดเลือกคุณลักษณะ	อัลกอริทึมจำแนก	การประเมินประสิทธิภาพ	ผลการวิจัย
จ	การเปรียบเทียบเทคนิคการสุ่มตัวอย่างเพื่อการจำแนกข้อมูลที่ไม่สมดุล	กาญจน์ ณ ศรีธะและคณะ	2561	SMOTE	ไม่มีการคัดเลือกคุณลักษณะ	- cart - random forest - support vector machine - neural network	- 10-fold cross validation - precision - recall - f-measure	เทคนิค adaboost ร่วมกับ random forest และเทคนิค bagging ร่วมกับ random forest
ฉ	การเปรียบเทียบการคัดเลือกคุณลักษณะที่สำคัญในการปรับปรุงการพยากรณ์มะเร็งเต้านม	อัจจิมา มณฑาพันธุ์	2562	-	- Correlation base Feature Selection - Information Gain - Gain Ratio - Chi Square - Forward Selection - Backward Elimination - Evolutionary Selection	support vector machine	- 10-fold cross validation - accuracy	วิธี Evolutionary Selection ร่วมกับเทคนิค support vector machine

ตารางที่ 2.4 ตารางเปรียบเทียบรายละเอียดงานวิจัยที่เกี่ยวข้อง (ต่อ)

ลำดับ	ชื่องานวิจัย	ผู้วิจัย	ปี	การจัดการข้อมูล	การคัดเลือกคุณลักษณะ	อัลกอริทึมจำแนก	การประเมินประสิทธิภาพ	ผลการวิจัย
ช	การลดจำนวนกลุ่มในการจำแนกแบบหลายกลุ่มเป็นสองกลุ่มสำหรับการจำแนกการกลับมารักษาซ้ำในโรงพยาบาลของผู้ป่วยโรคเบาหวาน	วิชญ์วิสิฐ เกษรสิทธิ์ และคณะ	2562	-	ไม่มีการคัดเลือกคุณลักษณะ	- decision tree - logistic regression	- accuracy	เทคนิค decision tree
ซ	การศึกษาปัจจัยที่ส่งผลต่อการสำเร็จการศึกษาตามแผนของนักศึกษาระดับปริญญาตรี โดยใช้เทคนิคการคัดเลือกคุณลักษณะบนชุดข้อมูลที่ไม่สมดุล	วรการ ใจดี และนพณัฐ วรรณภีร์ 2563	2564	SMOTE	- Correlation base Feature Selection - Information Gain - Chi Square	-	-	-

ตารางที่ 2.4 ตารางเปรียบเทียบรายละเอียดงานวิจัยที่เกี่ยวข้อง (ต่อ)

ลำดับ	ชื่องานวิจัย	ผู้วิจัย	ปี	การจัดการข้อมูล	การคัดเลือกคุณลักษณะ	อัลกอริทึมจำแนก	การประเมินประสิทธิภาพ	ผลการวิจัย
ฌ	An approach to overcome imbalance datasets of eukaryotic genomes during the analysis by Machine learning technique (svm).	Mohd.Faheem Khan, Gaurav Chauhan And A. K. Jaitly	2011	SMOTE	ไม่มีการคัดเลือกคุณลักษณะ	support vector machine	- f-measure - precision - recall	เทคนิค support vector machine
ญ	Feature selection and classification methods for decision making: a comparative analysis.	Osiris Villacampa	2015	-	- Information Gain - correlation based feature selection - relief-f	- decision tree - k-nearest neighbor - support vector machines	- 5-fold cross validation - accuracy - auc - tp rate & fp rate - f-measure	วิธี wrapper ร่วมกับเทคนิค vector machine

บทที่ 3
วิธีดำเนินการวิจัย

การวิจัยเรื่อง การเลือกคุณลักษณะสำคัญแบบพลวัตสำหรับความเหมาะสมที่สุดของการจำแนกต้นไม้ตัดสินใจบนพื้นฐานเงื่อนไขหลายเป้าหมาย นำเสนอการเลือกคุณลักษณะสำคัญแบบพลวัตสำหรับความเหมาะสมที่สุดกับการจำแนกต้นไม้ตัดสินใจบนพื้นฐานเงื่อนไขหลายเป้าหมาย โดยประยุกต์ใช้การเรียนรู้ของเครื่อง (Machine Learning : ML) เพื่อช่วยในการพยากรณ์การได้รับทุนการศึกษา ซึ่งงานวิจัยนี้มุ่งเน้นการเลือกคุณลักษณะสำคัญ เพื่อให้ได้คุณลักษณะสำคัญสำหรับนำไปสร้างแบบจำลองการพยากรณ์บนพื้นฐานเงื่อนไขหลายเป้าหมาย ผู้วิจัยได้ดำเนินตามขั้นตอนการดำเนินการวิจัย ดังนี้

1. ประชากรและกลุ่มตัวอย่าง
2. เครื่องมือที่ใช้ในการวิจัย
3. ขั้นตอนการดำเนินการวิจัย

1. ประชากรและกลุ่มตัวอย่าง

ข้อมูลนักศึกษาของวิทยาลัยเทคโนโลยีทางการแพทย์และสาธารณสุข กาญจนาภิเษก ซึ่งประกอบด้วยข้อมูลคุณลักษณะ จำนวน 29 คุณลักษณะ ข้อมูล จำนวน 500 แถว และเป้าหมายการได้รับทุนจำนวน 3 เป้าหมาย (จากฐานข้อมูลงานทะเบียนนักศึกษา และงานทุนการศึกษา วิทยาลัยเทคโนโลยีทางการแพทย์และสาธารณสุข กาญจนาภิเษก ณ วันที่ 8 กุมภาพันธ์ 2564)

2. เครื่องมือที่ใช้ในการวิจัย

ในการวิจัยนี้ผู้วิจัยได้กำหนดเครื่องมือที่ใช้ในการวิจัย ประกอบด้วย

2.1 เครื่องคอมพิวเตอร์

- ระบบปฏิบัติการ Windows 10 64 bit
- หน่วยความจำหลัก 4 GB

2.2 โปรแกรมจัดการตารางงาน (Microsoft Office Excel)

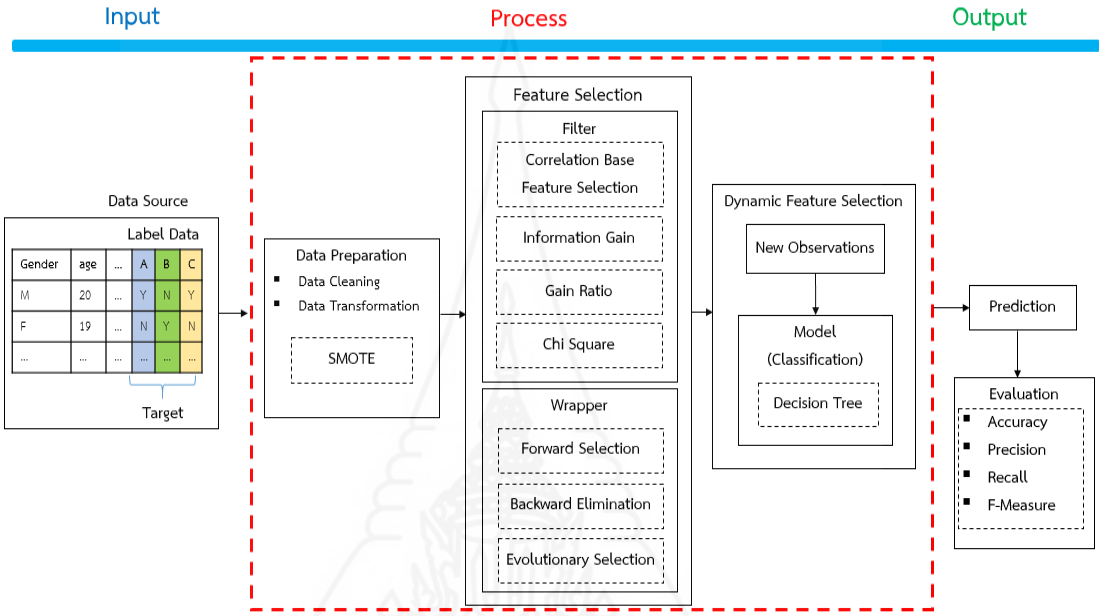
- สำหรับจัดเก็บข้อมูลและจัดเตรียมข้อมูลเพื่อนำไปใช้ในการวิเคราะห์ข้อมูล

2.3 โปรแกรมวิเคราะห์ข้อมูล (Rapid Miner)

- สำหรับออกแบบและการวิเคราะห์ข้อมูล

3. ขั้นตอนการดำเนินการวิจัย

การวิจัยนี้มีขั้นตอนการดำเนินการวิจัย 6 ขั้นตอนหลัก ได้แก่ 1) การเก็บรวบรวมข้อมูล (data collection) 2) การเตรียมข้อมูล (data preparation) 3) การสร้างแบบจำลอง (modeling) 4) การวัดประสิทธิภาพแบบจำลอง (evaluation the model) 5) การปรับค่าพารามิเตอร์ (parameter tuning) และ 6) การใช้งานการทำนายแบบจำลอง (model prediction) ดังภาพที่ 3.1



ภาพที่ 3.1 แสดงขั้นตอนการเลือกคุณลักษณะสำคัญแบบพลวัตสำหรับความเหมาะสมที่สุดของการจำแนกต้นไม้ตัดสินใจบนพื้นฐานเงื่อนไขหลายเป้าหมาย

จากภาพที่ 3.1 แสดงขั้นตอนการเลือกคุณลักษณะสำคัญแบบพลวัตสำหรับความเหมาะสมที่สุดของการจำแนกต้นไม้ตัดสินใจบนพื้นฐานเงื่อนไขหลายเป้าหมาย โดยขั้นตอนที่ 1 นำเข้าข้อมูลมาจากแหล่งข้อมูล (data source) ของข้อมูลนักศึกษาวิทยาลัยเทคโนโลยีทางการแพทย์และสาธารณสุข กาญจนาภิเษก ประกอบด้วยคุณลักษณะจำนวน 29 คุณลักษณะ แถวข้อมูลจำนวน 500 แถว และกำหนดเป้าหมายจำนวน 3 เป้าหมาย ได้แก่ class A, class B และ class C เป็นการกำหนดข้อมูลแบบมีป้ายกำกับ (labeled data) ขั้นตอนที่ 2 ทำเตรียมข้อมูลและการจัดการความไม่สมดุลของข้อมูล (imbalanced data classification) ด้วยวิธีวิธีการสุ่มตัวอย่างกลุ่มน้อย (SMOTE) เพื่อปรับสมดุลของข้อมูล ขั้นตอนที่ 3 นำข้อมูลเข้าสู่ขั้นตอนการวิเคราะห์และประมวลผล ด้วยเทคนิคการคัดเลือกคุณลักษณะสำคัญ โดยใช้วิธีการคัดเลือกคุณลักษณะ 2 วิธี ได้แก่ 1) วิธีแบบกรอง (Filter approach) ได้แก่ เทคนิค Correlation base Feature Selection เทคนิค Information Gain เทคนิค Gain Ratio และเทคนิค Chi Square และ 2) วิธีแบบควบรวม (Wrapper approach) ได้แก่ เทคนิค Forward Selection เทคนิค Backward Elimination และเทคนิค Evolutionary Selection เพื่อเลือกเทคนิคที่เหมาะสมบนพื้นฐานเงื่อนไขหลายเป้าหมาย ขั้นตอนที่ 4 เลือกคุณลักษณะสำคัญ

แบบพลวัต (dynamic feature selection) ซึ่งแปรผันไปตามเงื่อนไขหลายเป้าหมาย (multi-target conditions) ขั้นตอนที่ 5 สร้างแบบจำลองข้อมูลประเภทการจำแนกต้นไม้ตัดสินใจ (decision tree) และขั้นตอนที่ 6 ทำการประมินประสิทธิภาพแบบจำลอง โดยพิจารณาจากค่าความถูกต้อง (accuracy) ค่าความแม่นยำ (precision) ค่าเรียกคืน (recall) และค่าประสิทธิภาพโดยรวม (f-measure) รวมทั้งทำการทดสอบการปรับค่าพารามิเตอร์ต่าง ๆ เพื่อให้มีค่าที่มีความเหมาะสมกับชุดข้อมูล และแบบจำลองการจำแนกต้นไม้ตัดสินใจบนพื้นฐานเงื่อนไขหลายเป้าหมาย

3.1 การเก็บรวบรวมข้อมูล (data collection)

เก็บรวบรวมข้อมูลของนักศึกษาจากฐานข้อมูลงานทุนการศึกษาวิทยาลัยเทคโนโลยีทางการแพทย์และสาธารณสุข กาญจนาภิเษก ประกอบด้วยข้อมูล จำนวน 500 คน ข้อมูลปีการศึกษา 2562-2563 รายละเอียดดังนี้ 1) ข้อมูลนักศึกษา ปีการศึกษา 2562 จำนวน 251 คน และ 2) ข้อมูลนักศึกษา ปีการศึกษา 2563 จำนวน 249 คน โดยทำการเก็บข้อมูลประวัติของนักศึกษาที่ยื่นขอทุนซึ่งเป็นข้อมูลการได้รับทุนการศึกษาและไม่ได้รับทุนการศึกษา จากทุนทั้งหมด 3 ประเภท ได้แก่ ทุนได้เปล่า, ทุนกู้ยืมเพื่อการศึกษา และทุนขาดแคลนทุนทรัพย์ ซึ่งประกอบด้วยคุณลักษณะสำคัญจำนวน 29 คุณลักษณะ และเป้าหมาย จำนวน 3 เป้าหมาย ได้แก่ 1) เป้าหมาย A แทนทุนได้เปล่า 2) เป้าหมาย B แทนทุนกู้ยืมเพื่อการศึกษา และ 3) เป้าหมาย C แทนทุนขาดแคลนทุนทรัพย์

3.2 การเตรียมข้อมูล (data preparation)

ทำการเตรียมข้อมูลให้อยู่ในรูปแบบที่เหมาะสมสำหรับนำไปใช้ในการวิเคราะห์ จำนวน 500 แถว คุณลักษณะ จำนวน 29 คุณลักษณะสำคัญ และเป้าหมายจำนวน 3 เป้าหมาย โดยกำหนดการจำแนกประเภทข้อมูลของแต่ละเป้าหมายแบบไบนารี (binary) ได้แก่ “yes” แทนได้รับทุน และ “no” แทนไม่ได้รับทุน

3.2.1 การจัดการความไม่สมดุลของข้อมูล (imbalanced data classification)

ทำการปรับสมดุลของชุดข้อมูลเนื่องจากข้อมูลมีจำนวนการจำแนกประเภทแตกต่างกันจำนวนมาก คือ จำนวนการไม่ได้รับทุนมากกว่าการได้รับทุน ซึ่งมีความแตกต่างกันของคนที่ไม่ได้รับทุนกับคนที่ได้รับทุนแตกต่างกันมาก จึงใช้วิธีการสุ่มตัวอย่างกลุ่มน้อยสังเคราะห์ (Synthetic Minority Oversampling Technique: SMOTE) สำหรับเพิ่มข้อมูลให้มีจำนวนใกล้เคียงกัน เพื่อแก้ปัญหาการจำแนกข้อมูลเอนเอียงไปทางกลุ่มมากก่อนนำชุดข้อมูลไปทำการคัดเลือกคุณลักษณะสำคัญ

3.2.2 การเลือกคุณลักษณะสำคัญ (feature selection)

ใช้ชุดข้อมูลที่ผ่านการปรับสมดุลของข้อมูลด้วยวิธี SMOTE มาทำการคัดเลือกคุณลักษณะที่สำคัญ งานวิจัยนี้เลือกใช้วิธีการคัดเลือกคุณลักษณะ 2 วิธี ได้แก่ 1) วิธีแบบกรอง (Filter approach) ได้แก่ เทคนิค Correlation base Feature Selection เทคนิค Information Gain

เทคนิค Gain Ratio และเทคนิค Chi Square และ 2) วิธีแบบควบรวม (Wrapper approach) ได้แก่ เทคนิค Forward Selection เทคนิค Backward Elimination และเทคนิค Evolutionary Selection เพื่อเลือกเทคนิคที่เหมาะสมบนพื้นฐานเงื่อนไขหลายเป้าหมาย

3.2.3 การเลือกคุณลักษณะสำคัญแบบพลวัต (dynamic feature selection)

ทำการเลือกคุณลักษณะสำคัญแบบพลวัต ซึ่งคุณลักษณะสำคัญที่ได้จะแปรผันไปตามเป้าหมาย จำนวน 3 เป้าหมาย ได้แก่ 1) เป้าหมาย A แทนทุนได้เปล่า 2) เป้าหมาย B แทนทุนกู้ยืมเพื่อการศึกษา และ 3) เป้าหมาย C แทนทุนขาดแคลนทุนทรัพย์ โดยใช้เทคนิคการคัดเลือกคุณลักษณะที่หลากหลาย ทั้งหมด 7 เทคนิค จากนั้นพิจารณาเลือกเทคนิคที่เหมาะสมจากค่าความถูกต้อง (accuracy)

3.3 การสร้างแบบจำลอง (modeling)

สร้างแบบจำลองโดยนำข้อมูลทั้งหมด 500 แถว มาดำเนินการแบ่งข้อมูลออกเป็น 2 ส่วน ได้แก่ ข้อมูลฝึกสอน (training data) และข้อมูลทดสอบ (testing data) ออกเป็น 80:20 คือข้อมูลฝึกสอนคิดเป็นร้อยละ 80 หรือจำนวน 400 แถว และข้อมูลทดสอบคิดเป็นร้อยละ 20 หรือจำนวน 100 แถว จากนั้นนำข้อมูลมาสร้างแบบจำลองการจำแนกต้นไม้ตัดสินใจ และใช้วิธี 10-fold cross validation เพื่อให้ข้อมูลมีการกระจายค่าเท่าๆกัน

3.4 การวัดประสิทธิภาพโมเดล (evaluation the model)

วัดประสิทธิภาพแบบจำลอง โดยพิจารณาจากค่าความถูกต้อง (accuracy) ค่าความแม่นยำ (precision) ค่าเรียกคืน (recall) และค่าประสิทธิภาพโดยรวม (f-measure) ซึ่งเป็นการคำนวณจากตาราง confusion matrix

3.5 การปรับค่าพารามิเตอร์ (parameter tuning)

ทดสอบการปรับค่าพารามิเตอร์ต่าง ๆ เพื่อให้มีค่าที่มีความเหมาะสมกับชุดข้อมูล และแบบจำลองการจำแนกต้นไม้ตัดสินใจบนพื้นฐานเงื่อนไขหลายเป้าหมาย

3.6 การใช้งานการทำนายแบบจำลอง (model prediction)

การใช้งานการทำนายแบบจำลอง ถูกนำไปใช้งานจริงร่วมกับการพิจารณาของคณะกรรมการในการตัดสินใจให้ทุนแก่นักศึกษาของวิทยาลัยฯ อย่างยุติธรรม และโปร่งใส

บทที่ 4
ผลการดำเนินการวิจัย

การวิจัยเรื่อง การเลือกคุณลักษณะสำคัญแบบพลวัตสำหรับความเหมาะสมที่สุดของการจำแนกต้นไม้ตัดสินใจบนพื้นฐานเงื่อนไขหลายเป้าหมาย ผู้วิจัยการดำเนินประกอบด้วย 6 ขั้นตอนหลัก ได้แก่ 1) การเก็บรวบรวมข้อมูลของนักศึกษาจากงานทุนการศึกษาของวิทยาลัยเทคโนโลยีทางการแพทย์และสาธารณสุข กาญจนาภิเษก ข้อมูลจำนวน 500 คน และคุณลักษณะสำคัญจำนวน 29 คุณลักษณะ, 2) การเตรียมข้อมูลโดยกำหนดเงื่อนไขหลายเป้าหมายจำนวน 3 แบบ สำหรับทุนที่มี 3 ประเภท โดยประยุกต์วิธีสังเคราะห์ข้อมูลเพิ่มสำหรับแก้ไขปัญหาชุดข้อมูลของนักศึกษาที่มีความไม่สมดุล และพัฒนาวิธีการเลือกคุณลักษณะสำคัญแบบพลวัตบนพื้นฐานเงื่อนไขหลายเป้าหมาย, 3) การสร้างแบบจำลองการจำแนกด้วยอัลกอริทึมต้นไม้ตัดสินใจ สำหรับสอนและทดสอบแบบจำลอง, 4) การประเมินประสิทธิภาพแบบจำลอง, 5) การปรับค่าพารามิเตอร์เพื่อหาค่าความเหมาะสมที่สุด และ 6) การใช้งานการทำนายแบบจำลอง ได้ผลการดำเนินการวิจัยดังนี้

1. การเก็บรวบรวมข้อมูล (data collection)
2. การเตรียมข้อมูล (data preparation)
3. การสร้างแบบจำลอง (modeling)
4. การวัดประสิทธิภาพโมเดล (evaluation the model)
5. การปรับค่าพารามิเตอร์ (parameter tuning)
6. การใช้งานการทำนายแบบจำลอง (model prediction)

1. การเก็บรวบรวมข้อมูล (data collection)

ทำการเก็บรวบรวมข้อมูลของนักศึกษาจากฐานข้อมูลงานทุนการศึกษาวิทยาลัยเทคโนโลยีทางการแพทย์และสาธารณสุข กาญจนาภิเษก ประกอบด้วยข้อมูล จำนวน 500 คน ข้อมูลปีการศึกษา 2562-2563 รายละเอียดดังนี้ 1) ปีการศึกษา 2562 จำนวน 251 คน และปีการศึกษา 2563 จำนวน 249 คน ซึ่งประกอบด้วยคุณลักษณะสำคัญจำนวน 29 คุณลักษณะ และเป้าหมายจำนวน 3 เป้าหมาย ได้แก่ 1) เป้าหมาย A แทนทุนได้เปล่า 2) เป้าหมาย B แทนทุนกู้ยืมเพื่อการศึกษา และ 3) เป้าหมาย C แทนทุนขาดแคลนทุนทรัพย์ รายละเอียดคุณลักษณะดังตารางที่ 4.1

ตารางที่ 4.1 แสดงรายละเอียดคุณลักษณะของข้อมูล

ลำดับ	แอตทริบิวต์	รายละเอียด
1	Gender	เพศ
2	Age	อายุ
3	Nationality	สัญชาติ
4	Race	เชื้อชาติ
5	Religion	ศาสนา
6	Blood	หมู่เลือด
7	Disease	โรคประจำตัว
8	Disability	ความพิการ
9	Talent	ความสามารถพิเศษ
10	Quota	โควต้าทุน
11	Hometown	ภูมิลำเนา
12	address	ที่อยู่ตามภูมิลำเนา
13	Famsize	จำนวนพี่น้อง
14	Son_number	บุตรลำดับที่
15	F_Status	สถานภาพบิดา
16	F_Occup	ประเภทอาชีพบิดา
17	F_Income	รายได้บิดา
18	M_status	สถานภาพมารดา
19	M_Occup	ประเภทอาชีพมารดา
20	M_Income	รายได้มารดา
21	P_status	สถานภาพบิดามารดา
22	Guardian	ผู้ปกครอง
23	G_Occup	ประเภทอาชีพผู้ปกครอง
24	G_Income	รายได้ผู้ปกครอง
25	Old_Edu	ระดับการศึกษาเดิม
26	Old_Gpax	เกรดเฉลี่ยสะสมเดิม
27	Recruit	วิธีสมัคร
28	Edu_Level	ระดับการศึกษาปัจจุบัน
29	Major	สาขาที่เรียน
30	**Target A**	**ทุนได้เปล่า**
31	**Target B**	**ทุนกู้ยืมเพื่อการศึกษา**
32	**Target C**	**ทุนขาดแคลนทุนทรัพย์**

จากตารางที่ 4.1 เป็นการเตรียมชุดข้อมูลที่ประกอบด้วยคุณลักษณะสำคัญ 29 คุณลักษณะ และการกำหนดเป้าหมายผลลัพธ์ที่ต้องการไว้เป็นแอตทริบิวต์ 3 ลำดับสุดท้าย ให้อยู่ในรูปแบบที่พร้อมสำหรับนำไปใช้ในการวิเคราะห์ตามหลักการเรียนรู้ของเครื่อง (machine learning)

2. การเตรียมข้อมูล (data preparation)

ทำการเตรียมข้อมูลให้อยู่ในรูปแบบที่เหมาะสมสำหรับนำไปใช้ในการวิเคราะห์ จำนวน 500 แถว คุณลักษณะ จำนวน 29 คุณลักษณะสำคัญ และเป้าหมายจำนวน 3 เป้าหมาย โดยกำหนดการจำแนกประเภทข้อมูลของแต่ละเป้าหมายแบบไบนารี (binary) ได้แก่ “yes” แทนได้รับทุน และ “no” แทนไม่ได้รับทุน จากนั้นนำชุดข้อมูลเข้าสู่โปรแกรม rapid miner ซึ่งผู้วิจัยดำเนินการดังนี้

2.1 การกลั่นกรองข้อมูล (data cleaning)

การกลั่นกรองข้อมูล (data cleaning) เป็นการตรวจสอบความถูกต้องและความสมบูรณ์ของข้อมูล การจัดการกับข้อมูลที่ขาดหายด้วยการเติมเต็มข้อมูล และการจัดการข้อมูลที่มีความผิดพลาด ผิดปกติ และข้อมูลไม่มีความสอดคล้องกัน โดยการจัดการข้อมูลให้มีความถูกต้องสมบูรณ์เพื่อใช้ในการวิเคราะห์ข้อมูล ซึ่งในขั้นตอนนี้อาจพบลักษณะของข้อมูลที่ไม่เหมาะสม ได้แก่ ข้อมูลไม่สมบูรณ์ (incomplete data) เช่น ข้อมูลขาดหาย (missing value) ข้อมูลรบกวน (noisy data) และข้อมูลไม่สอดคล้อง (inconsistent data) ซึ่งงานวิจัยนี้พบลักษณะข้อมูลที่ไม่เหมาะสมดังตัวอย่าง ภาพที่ 4.1 – ภาพที่ 4.2

สถานภาพบิดา	ประเภทอาชีพบิดา	รายได้ต่อเดื	สถานภาพมารดา	ประเภทอาชีพมารดา	รายได้ต่อเดื	สถานภาพของบิด
ยังมีชีวิตอยู่	เกษตรกร/ประมง	8,333	ยังมีชีวิตอยู่	ข้าราชการ/เจ้าหน้าที่		
ยังมีชีวิตอยู่	อื่นๆ	14,000	ยังมีชีวิตอยู่	อื่นๆ		
ยังมีชีวิตอยู่	อาชีพอิสระ/ธุรกิจอิสระ/	25,000	ยังมีชีวิตอยู่	บริษัท/องค์กรธุรกิจเ		
ยังมีชีวิตอยู่	อื่นๆ	8,000	ยังมีชีวิตอยู่	อื่นๆ	6,667	สมรส
ยังมีชีวิตอยู่	ข้าราชการ/เจ้าหน้าที่หน่	16,667	ยังมีชีวิตอยู่	เกษตรกร/ประมง	8,333	สมรส
ยังมีชีวิตอยู่	เกษตรกร/ประมง	2,500	ถึงแก่กรรม	เกษตรกร/ประมง	4,167	สมรส
ยังมีชีวิตอยู่	เกษตรกร/ประมง	833	ยังมีชีวิตอยู่	อื่นๆ	2,500	สมรส
ยังมีชีวิตอยู่	อื่นๆ	5,000	ยังมีชีวิตอยู่	อื่นๆ	4,167	สมรส
ยังมีชีวิตอยู่	ไม่มี	ไม่มี	ยังมีชีวิตอยู่	อื่นๆ	ไม่มี	แยกกันอยู่
ยังมีชีวิตอยู่	อื่นๆ	8,333	ยังมีชีวิตอยู่	อื่นๆ	4,167	สมรส
ยังมีชีวิตอยู่	เกษตรกร/ประมง	6,667	ยังมีชีวิตอยู่	เกษตรกร/ประมง	7,500	สมรส
ยังมีชีวิตอยู่	บริษัท/องค์กรธุรกิจเอกช	6,600	ยังมีชีวิตอยู่	อื่นๆ	8,058	สมรส
ยังมีชีวิตอยู่	อื่นๆ	11,667	ยังมีชีวิตอยู่	อื่นๆ	11,667	สมรส

ข้อมูลที่ไม่สอดคล้อง

ภาพที่ 4.1 แสดงตัวอย่างข้อมูลที่ไม่สอดคล้อง (inconsistent data) และข้อมูลที่มีค่าผิดพลาด (error)

จากภาพที่ 4.1 พบว่า มีข้อมูลที่มีค่าไม่สอดคล้อง (inconsistent data) ซึ่งมีค่าเกินความเป็นจริง แก้ไขโดยพิจารณาจากแก้ไขข้อมูลที่มีความขัดแย้ง คือ ปรับค่าของข้อมูลในส่วนของสถานภาพมารดาซึ่งไม่สอดคล้องกับข้อมูลประเภทอาชีพมารดา และรายได้ต่อเดือน ได้พิจารณาการปรับค่าให้มีความสอดคล้อง คือ สถานภาพมารดาจากเดิมมีค่าเป็น “ถึงแก่กรรม” เปลี่ยนเป็น “ยังมีชีวิตอยู่” เป็นต้น

สถานภาพมารดา	ประเภทอาชีพมารดา	รายได้ต่อเดือ	สถานภาพของบิดา	สถานภาพบิดามาร	ผู้ปกครอง	ประเภทอาชีพผู้ปกค
ยังมีชีวิตอยู่	ข้าราชการ/เจ้าหน้าที่หน่วยงานรั	45,000	สมรส	อยู่ด้วยกัน	มารดา	ข้าราชการ/เจ้าหน้า
ยังมีชีวิตอยู่	อื่นๆ	3,000	สมรส	อยู่ด้วยกัน	มารดา	อื่นๆ
ยังมีชีวิตอยู่	บริษัท/องค์กรธุรกิจเอกชน	22,000	อย		ดา	บริษัท/องค์กรธุรกิจเ
ยังมีชีวิตอยู่	อื่นๆ	6,667	ส		ดา	อื่นๆ
ยังมีชีวิตอยู่	เกษตรกร/ประมง	8,333	สมรส	อยู่ด้วยกัน	มารดา	เกษตรกร/ประมง
ยังมีชีวิตอยู่	เกษตรกร/ประมง	4,167	สม	อยู่ด้วยกัน	มารดา	เกษตรกร/ประมง
ยังมีชีวิตอยู่	อื่นๆ	2,500		อยู่ด้วยกัน	มารดา	อื่นๆ
ยังมีชีวิตอยู่	อื่นๆ	4,167	สมรส	อยู่ด้วยกัน	บิดา	อื่นๆ
ยังมีชีวิตอยู่	อื่นๆ	12300 บาท	แยกกันอยู่	แยกกันอยู่	อื่นๆ	เกษตรกร/ประมง
ยังมีชีวิตอยู่	อื่นๆ	4,167	สมรส	อยู่ด้วยกัน	บิดา	อื่นๆ
ยังมีชีวิตอยู่	เกษตรกร/ประมง	7,500	สมรส	อยู่ด้วยกัน	มารดา	เกษตรกร/ประมง

ค่าที่ผิดพลาด

ภาพที่ 4.2 แสดงตัวอย่างข้อมูลที่มีค่าผิดพลาด (error)

จากภาพที่ 4.2 พบว่า มีข้อมูลที่มีค่าผิดพลาด (error) ที่มีการรบกวน (noisy data) คือ ข้อมูลที่ผิดพลาดจากการกรอกข้อมูลผ่านระบบโดยผู้ใช้งาน (user) แก้ไขโดยพิจารณาจากแก้ไขข้อมูลที่มีความขัดแย้ง คือ ปรับค่าข้อมูลรายได้ต่อเดือนของบิดา จากเดิม “12300 บาท” เปลี่ยนเป็น “12,300”เป็นต้น

1.1 การแปลงรูปข้อมูล (data transformation)

เป็นการเตรียมข้อมูลให้อยู่ในรูปแบบที่พร้อมสำหรับนำไปใช้ในการวิเคราะห์ ตามหลักการเรียนรู้ของเครื่อง (machine learning) ดังตารางที่ 4.2

ตารางที่ 4.2 แสดงการแปลงรูปข้อมูลนักศึกษาให้อยู่ในรูปแบบที่พร้อมสำหรับนำไปใช้วิเคราะห์ข้อมูล

No	Detail	Attribute	Value	Code	Description
1	เพศ	Gender	nominal	F	หญิง
				M	ชาย
2	อายุ	Age	ordinal	A1	18-20 ปี
				A2	21-25 ปี
				A3	26-30 ปี
				A4	31 ปี ขึ้นไป
3	สัญชาติ	Nationality	nominal	Thai	ไทย
				Other	อื่นๆ
4	เชื้อชาติ	Race	nominal	Thai	ไทย
				Other	อื่นๆ
5	ศาสนา	Religion	nominal	Buddhism	พุทธ
				Christian	คริสต์
				Islam	อิสลาม
				Other	อื่นๆ

ตารางที่ 4.2 แสดงการแปลงรูปข้อมูลนักศึกษาให้อยู่ในรูปแบบที่พร้อมสำหรับนำไปใช้วิเคราะห์ข้อมูล (ต่อ)

No	Detail	Attribute	Value	Code	Description
6	หมู่เลือด	Blood	nominal	A	a
				B	b
				AB	ab
				O	o
7	โรคประจำตัว	Disease	nominal	Yes	มี
				No	ไม่มี
8	ความพิการ	Disability	nominal	Yes	มี
				No	ไม่มี
9	ความสามาถ พิเศษ	Talent	nominal	Yes	มี
				No	ไม่มี
10	โควต้าทุน	Quota	nominal	Health	สาธารณสุขจังหวัด
				Hos	โรงพยาบาล
				edu	ทุนวิทยาลัย
				Other	อื่นๆ
11	ภูมิลำเนา	Hometown	nominal	Bangkok	กรุงเทพ
				Central	ภาคกลาง
				North	ภาคเหนือ
				South	ภาคใต้
				Northeast	ภาคตะวันออกเฉียงเหนือ
12	ประเภทที่อยู่ ตามภูมิลำเนา	address	nominal	City	ในเมือง
				Country	นอกเมือง
13	มีจำนวนพี่น้องร่วมบิดามารดา	Famsize	numeric		จำนวนพี่น้อง (คน)
14	เป็นบุตรลำดับที่	Son_number	numeric		ลำดับที่
15	สถานภาพ บิดา	F_Status	nominal	Fs1	ยังมีชีวิตอยู่
				Fs2	ถึงแก่กรรม
16	ประเภทอาชีพ บิดา	F_Occup	nominal	Gover	ข้าราชการ/เจ้าหน้าที่หน่วยงานรัฐ
				State	รัฐวิสาหกิจ
				Compa	บริษัท/องค์กรธุรกิจเอกชน
				Free	อาชีพอิสระ/ธุรกิจอิสระ/เจ้าของกิจการ
				Far_Fish	เกษตรกร/ประมง
				other	อื่น ๆ

ตารางที่ 4.2 แสดงการแปลงรูปข้อมูลนักศึกษาให้อยู่ในรูปแบบที่พร้อมสำหรับนำไปใช้วิเคราะห์ข้อมูล (ต่อ)

No	Detail	Attribute	Value	Code	Description
17	รายได้ต่อเดือน	F_Income	ordinal	None	ไม่มีรายได้
				Lower	รายได้น้อยกว่าเท่ากับ 10,000 บาท
				Low	รายได้ 10,001 - 15,000 บาท
				Medium	รายได้ 15,001 - 20,000 บาท
				High	รายได้ 20,001 - 25,000 บาท
				Higher	รายได้มากกว่า 25,000 บาท
18	สถานภาพมารดา	M_status	nominal	Ms1	ยังมีชีวิตอยู่
				Ms2	ถึงแก่กรรม
19	ประเภทอาชีพมารดา	M_Occup	nominal	Gover	ข้าราชการ/เจ้าหน้าที่หน่วยงานรัฐ
				State	รัฐวิสาหกิจ
				Compa	บริษัท/องค์กรธุรกิจเอกชน
				Free	อาชีพอิสระ/ธุรกิจอิสระ/เจ้าของกิจการ
				Far_Fish	เกษตรกร/ประมง
				other	อื่น ๆ
20	รายได้ต่อเดือน	M_Income	ordinal	None	ไม่มีรายได้
				Lower	รายได้น้อยกว่าเท่ากับ 10,000 บาท
				Low	รายได้ 10,001 - 15,000 บาท
				Medium	รายได้ 15,001 - 20,000 บาท
				High	รายได้ 20,001 - 25,000 บาท
				Higher	รายได้มากกว่า 25,000 บาท
21	สถานภาพบิดามารดา	P_status	nominal	Together	อยู่ด้วยกัน
				Divorce	หย่าร้าง
				Separate	แยกกันอยู่
				Other	อื่น ๆ
22	ผู้ปกครอง	Guardian	nominal	Family	บิดาและมารดา
				Father	บิดา
				Mother	มารดา
				Relative	อื่น ๆ

ตารางที่ 4.2 แสดงการแปลงรูปข้อมูลนักศึกษาให้อยู่ในรูปแบบที่พร้อมสำหรับนำไปใช้วิเคราะห์ข้อมูล (ต่อ)

No	Detail	Attribute	Value	Code	Description
23	ประเภทอาชีพผู้ปกครอง	G_Occup	nominal	Gover	ข้าราชการ/เจ้าหน้าที่หน่วยงานรัฐ
				State	รัฐวิสาหกิจ
				Compa	บริษัท/องค์กรธุรกิจเอกชน
				Free	อาชีพอิสระ/ธุรกิจอิสระ/เจ้าของกิจการ
				Far_Fish	เกษตรกร/ประมง
				other	อื่น ๆ
24	รายได้ผู้ปกครองต่อเดือน	G_Income	ordinal	Non	ไม่มีรายได้
				Lower	รายได้น้อยกว่าเท่ากับ 10,000 บาท
				Low	รายได้ 10,001 - 15,000 บาท
				Medium	รายได้ 15,001 - 20,000 บาท
				High	รายได้ 20,001 - 25,000 บาท
				Higher	รายได้มากกว่า 25,000 บาท
25	ระดับการศึกษาเดิม	Old_Edu	nominal	Level1	ระดับมัธยมศึกษาตอนปลาย
				Level2	ระดับอนุปริญญา
				Level3	ระดับปริญญาตรี
				Level4	ระดับที่สูงกว่าปริญญาตรี
				Other	อื่น ๆ
26	เกรดเฉลี่ยสะสมเดิม	Old_Gpax	ordinal	Medium	เฉลี่ย 2.00-2.49
				Good	เฉลี่ย 2.50-2.99
				Excellent	เฉลี่ย 3.00-3.49
				Best	เฉลี่ย 3.50 ขึ้นไป
27	วิธีสมัคร	Recruit	nominal	R1	รอบ admissions
				R2	รอบการรับตรงอิสระ
				R3	รอบรับตรงจากพื้นที่
				R4	อื่น ๆ
28	ปัจจุบันศึกษาในระดับ	Edu_Level	nominal	EL1	ระดับประกาศนียบัตรวิชาชีพชั้นสูง
				EL2	ระดับปริญญาตรี
29	สาขาที่เรียน	Major	nominal	mrs	เวชระเบียน
				ttm	การแพทย์แผนไทย
				av	โสตทัศนศึกษาทางการแพทย์

ตารางที่ 4.2 แสดงการแปลงรูปข้อมูลนักศึกษาให้อยู่ในรูปแบบที่พร้อมสำหรับนำไปใช้วิเคราะห์ข้อมูล (ต่อ)

No	Detail	Attribute	Value	Code	Description
30	ทุนได้เปล่า	Target A	nominal	Yes	ได้รับทุน
				No	ไม่ได้รับทุน
31	ทุนกู้ยืมเพื่อการศึกษา	Target B	nominal	Yes	ได้รับทุน
				No	ไม่ได้รับทุน
32	ทุนขาดแคลนทุนทรัพย์	Target C	nominal	Yes	ได้รับทุน
				No	ไม่ได้รับทุน

จากภาพที่ 4.2 แสดงรายละเอียดคุณลักษณะข้อมูลของชุดข้อมูลทั้งหมดมีคุณลักษณะ ประกอบด้วยคุณลักษณะสำคัญจำนวน 29 คุณลักษณะ และเป้าหมาย จำนวน 3 เป้าหมาย ได้แก่ 1) เป้าหมาย A แทนทุนได้เปล่า 2) เป้าหมาย B แทนทุนกู้ยืมเพื่อการศึกษา และ 3) เป้าหมาย C แทนทุนขาดแคลนทุนทรัพย์ ที่จะใช้สำหรับนำเข้าสู่โปรแกรมเพื่อวิเคราะห์ข้อมูล

2.1 การจัดการความไม่สมดุลของข้อมูล (imbalanced data classification)

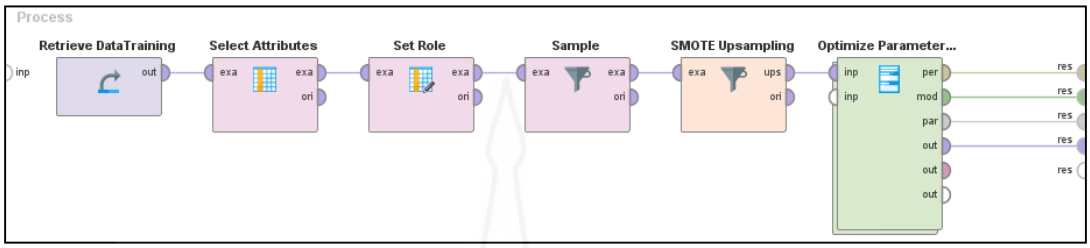
ทำการปรับสมดุลของชุดข้อมูลเนื่องจากข้อมูลมีจำนวนการจำแนกประเภทแตกต่างกันจำนวนมาก คือ จำนวนการไม่ได้รับทุนมากกว่าการได้รับทุน ซึ่งมีความแตกต่างกันของคนที่ไม่ได้รับทุนกับคนที่ได้รับทุนแตกต่างกันมาก ดังตารางที่ 4.3

ตารางที่ 4.3 แสดงข้อมูลที่มีความไม่สมดุล (imbalanced data)

Target	ได้รับทุน (yes)	ไม่ได้รับทุน (no)
A	102	398
B	460	40
C	94	406

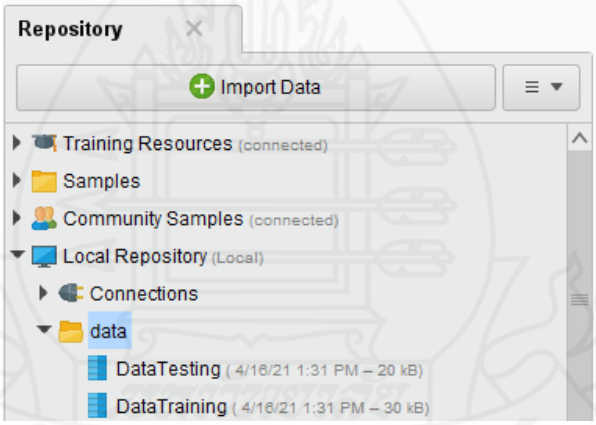
จากตารางที่ 4.3 แสดงจำนวนข้อมูลการจำแนกประเภทที่มีความแตกต่างกันจำนวนมาก ซึ่งมีความแตกต่างกันของข้อมูลการได้รับทุนประเภทต่างๆ คือ คนที่ไม่ได้รับทุนกับคนที่ได้รับทุนแตกต่างกันจำนวนมาก

ผู้วิจัยจึงใช้วิธีการสุ่มตัวอย่างกลุ่มน้อยสังเคราะห์ (Synthetic Minority Oversampling Technique: SMOTE) สำหรับเพิ่มข้อมูลให้มีจำนวนใกล้เคียงกัน เพื่อแก้ปัญหาการจำแนกข้อมูลเอนเอียงไปทางกลุ่มมากก่อนนำชุดข้อมูลไปทำการคัดเลือกคุณลักษณะสำคัญ ภาพรวมกระบวนการดังภาพที่ 4.3

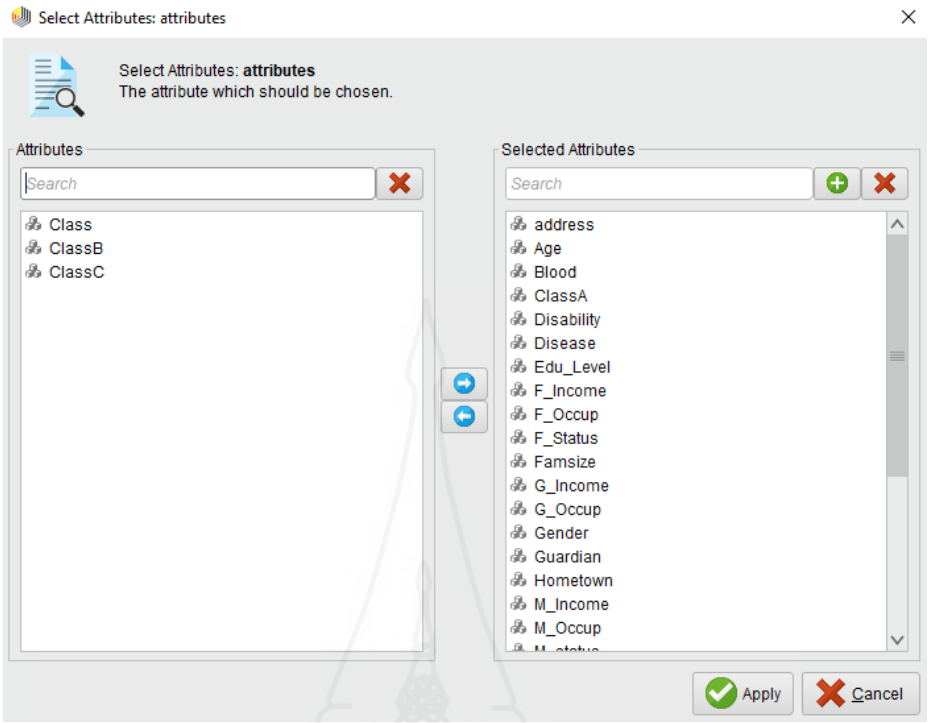


ภาพที่ 4.3 แสดงภาพรวมกระบวนการนำเข้าข้อมูลและการปรับสมดุลของข้อมูลด้วยวิธี SMOTE

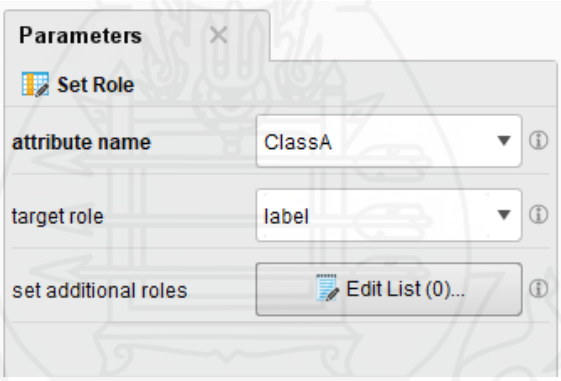
จากภาพที่ 4.3 แสดงกระบวนการนำเข้าข้อมูล (training data) โดยใช้ retrieve ดังภาพที่ 4.4 ทำการเลือกคุณลักษณะของข้อมูลโดยใช้ select attributes ดังภาพที่ 4.5 และกำหนดเป้าหมาย โดยใช้ set role ดังภาพที่ 4.6 ใช้สำหรับปรับเปลี่ยนเป้าหมายที่ต้องการ จากนั้นทำการปรับสมดุลข้อมูลด้วยวิธี SMOTE ดังภาพที่ 4.7 – 4.8 และกำหนด optimize parameters (grid) เพื่อกำหนดค่าพารามิเตอร์ที่เหมาะสม



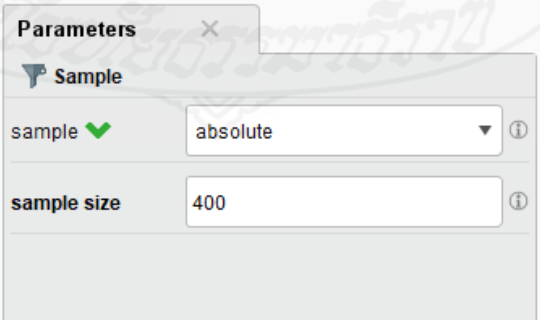
ภาพที่ 4.4 แสดงการนำเข้าข้อมูล (training data)



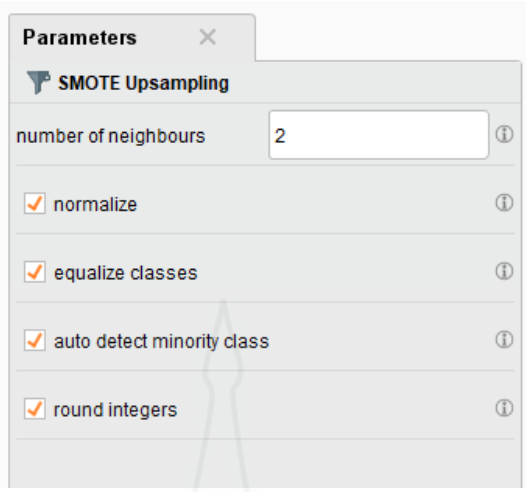
ภาพที่ 4.5 แสดงการเลือกคุณลักษณะของข้อมูล



ภาพที่ 4.6 แสดงการกำหนดเป้าหมาย



ภาพที่ 4.7 แสดงการกำหนดจำนวนข้อมูล sample



ภาพที่ 4.8 แสดงการปรับสมดุลข้อมูลด้วยวิธี SMOTE

เมื่อทำการจัดการความไม่สมดุลของข้อมูล ด้วยวิธีการสุ่มตัวอย่างกลุ่มน้อยสังเคราะห์ (Synthetic Minority Oversampling Technique: SMOTE) สำหรับเพิ่มข้อมูลให้มีจำนวนใกล้เคียงกัน ได้ผลการทดลองดังตารางที่ 4.4

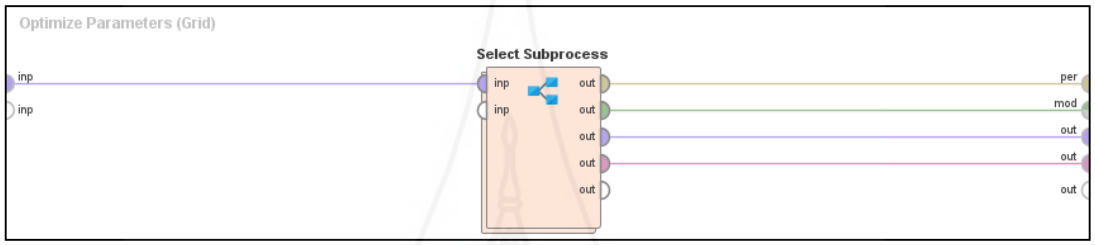
ตารางที่ 4.4 แสดงข้อมูลที่ผ่านการปรับสมดุลข้อมูลด้วยวิธี SMOTE

target	original		SMOTE	
	ได้รับทุน (yes)	ไม่ได้รับทุน (no)	ได้รับทุน (yes)	ไม่ได้รับทุน (no)
A	102	398	398	398
B	460	40	460	460
C	94	406	406	406

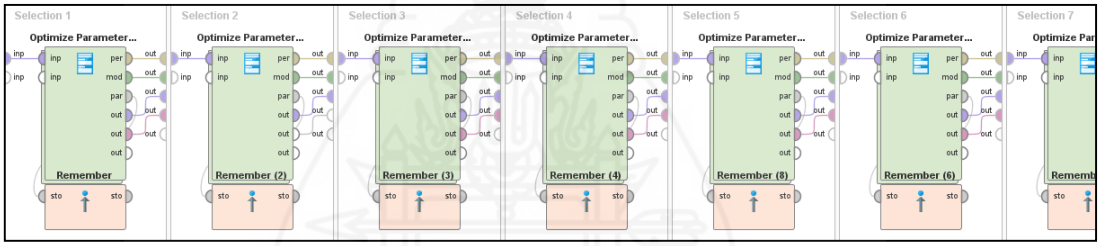
จากตารางที่ 4.4 แสดงข้อมูลที่ผ่านการปรับสมดุลของข้อมูลด้วยวิธี SMOTE ซึ่งเป็นการสุ่มปรับเพิ่มตัวอย่างกลุ่มน้อย

2.2 การเลือกคุณลักษณะสำคัญ (feature selection)

ใช้ชุดข้อมูลที่ผ่านการปรับสมดุลของข้อมูลด้วยวิธี SMOTE มาทำการคัดเลือกคุณลักษณะที่สำคัญ งานวิจัยนี้เลือกใช้วิธีการคัดเลือกคุณลักษณะ 2 วิธี ได้แก่ 1) วิธีแบบกรอง (filter approach) ได้แก่ เทคนิค Correlation base Feature Selection เทคนิค Information Gain เทคนิค Gain Ratio และเทคนิค Chi Square และ 2) วิธีแบบควบรวม (wrapper approach) ได้แก่ เทคนิค Forward Selection เทคนิค Backward Elimination และเทคนิค Evolutionary Selection เพื่อเปรียบเทียบ และเลือกเทคนิคที่เหมาะสมบนพื้นฐานเงื่อนไขหลายเป้าหมาย ภาพรวมกระบวนการดังภาพที่ 4.9



(ก)



(ข)

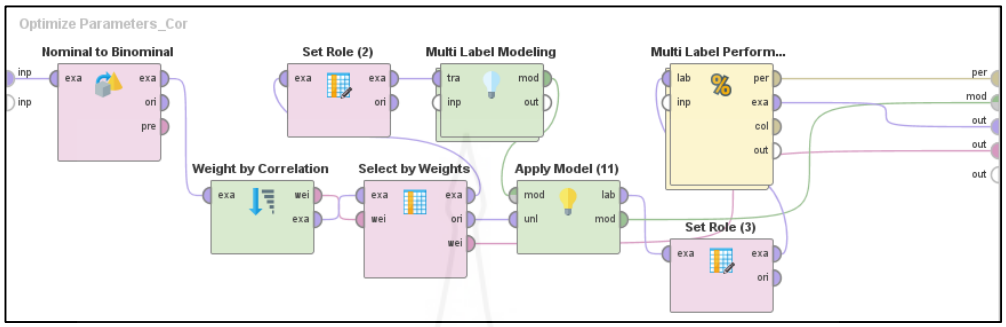
ภาพที่ 4.9 แสดงภาพรวมกระบวนการเลือกเทคนิคที่เหมาะสมสำหรับการเลือกคุณลักษณะสำคัญแบบพลวัต

จากภาพที่ 4.9 แสดงภาพกระบวนการเลือกเทคนิคที่เหมาะสมสำหรับการเลือกคุณลักษณะสำคัญแบบพลวัต ภาพที่ 4.2 (ก) ในส่วนของ input เป็นการนำเข้าข้อมูลที่ผ่านกระบวนการปรับสมดุลของข้อมูลด้วยวิธี SMOTE (จากภาพที่ 4.1) จากนั้นกำหนด select subprocess ให้คัดเลือกเทคนิคที่เหมาะสมที่สุดมาเพียง 1 เทคนิค โดยคัดเลือกจาก selection1 - selection7 และกำหนดให้แสดงผลลัพธ์ output เพื่อแสดงผลลัพธ์ค่า performance, model, exampleset และ weights และค่าอื่น ๆ ที่ต้องการแสดงของเทคนิคที่คัดเลือกได้จากทั้งหมด 7 เทคนิค ภาพที่ 4.2(ข)

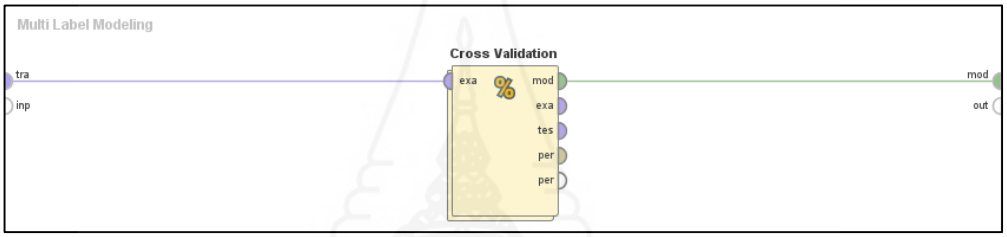
จากภาพที่ 4.9(ข) เป็นการกำหนด selection1 - selection7 ของเทคนิคการเลือกคุณลักษณะสำคัญของทั้งหมด 7 เทคนิค ซึ่งเป็นการเปรียบเทียบประสิทธิภาพการคัดเลือกคุณลักษณะสำคัญร่วมกับการจำแนกต้นไม้ตัดสินใจ ของการคัดเลือกคุณลักษณะสำคัญของทั้ง 7 เทคนิค รายละเอียดแสดงดังภาพที่ 4.10 – 4.16

2.2.1 เทคนิค Correlation base Feature Selection

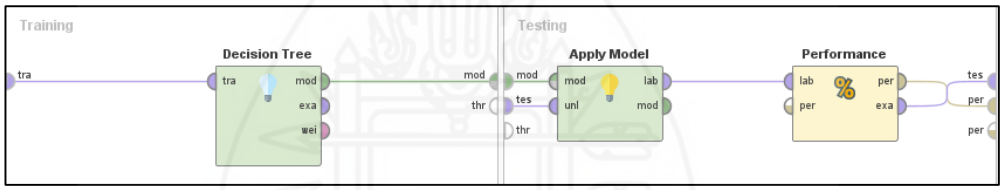
ผู้วิจัยทำการออกแบบการคัดเลือกคุณลักษณะสำคัญของเทคนิค Correlation base Feature Selection ใน selection1 ดังภาพที่ 4.10



(ก)



(ข)



(ค)

ภาพที่ 4.10 แสดงการออกแบบการคัดเลือกคุณลักษณะสำคัญของเทคนิค Correlation base Feature Selection

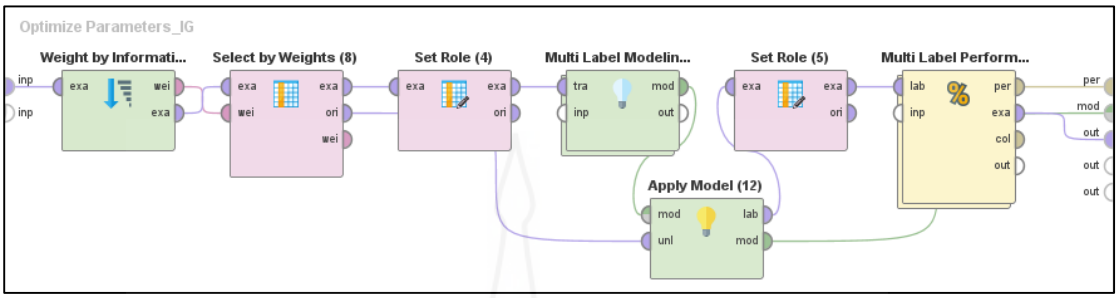
จากภาพที่ 4.10 แสดงการออกแบบการคัดเลือกคุณลักษณะสำคัญของเทคนิค Correlation base Feature Selection ของการจำแนกประเภทข้อมูลด้วยต้นไม้ตัดสินใจบนพื้นฐานเงื่อนไขหลายเป้าหมาย เพื่อทำการหาค่าความถูกต้องของเทคนิคการคัดเลือกคุณลักษณะสำคัญ ได้ผลการทดลองดังตารางที่ 4.5

ตารางที่ 4.5 แสดงค่าความถูกต้องของเทคนิค Correlation base Feature Selection

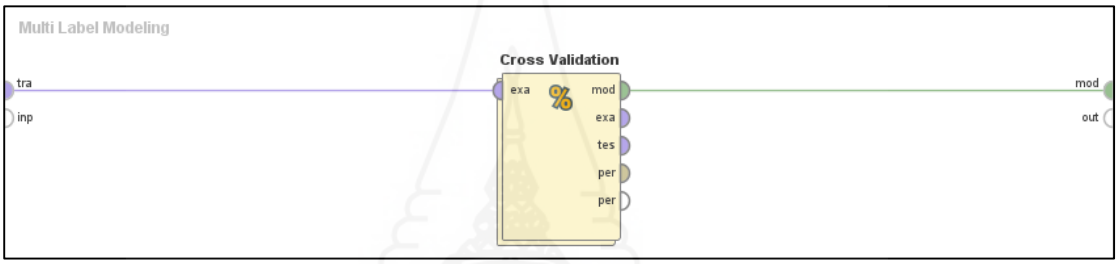
Target	Accuracy
A	0.863
B	0.915
C	0.993
Average	0.923

2.2.2 เทคนิค Information Gain

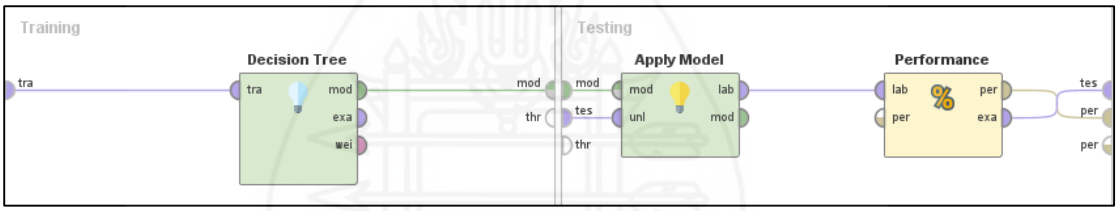
ผู้วิจัยทำการออกแบบการคัดเลือกคุณลักษณะสำคัญของเทคนิค Information Gain ใน selection2 ดังภาพที่ 4.11



(ก)



(ข)



(ค)

ภาพที่ 4.11 แสดงการออกแบบการคัดเลือกคุณลักษณะสำคัญของเทคนิค

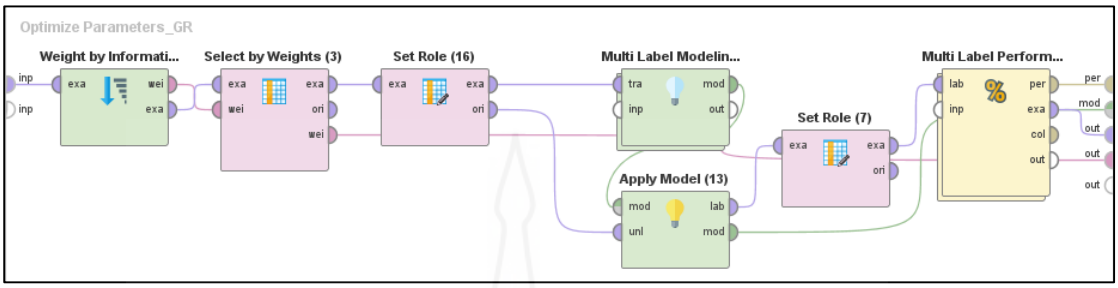
จากภาพที่ 4.11 แสดงการออกแบบการคัดเลือกคุณลักษณะสำคัญของเทคนิค Information Gain ของการจำแนกประเภทข้อมูลด้วยต้นไม้ตัดสินใจบนพื้นฐานเงื่อนไขหลายเป้าหมาย เพื่อทำการหาค่าความถูกต้องของเทคนิคการคัดเลือกคุณลักษณะสำคัญ สำหรับคัดเลือกเทคนิคการเลือกคุณลักษณะสำคัญที่เหมาะสมบนพื้นฐานเงื่อนไขหลายเป้าหมาย ได้ผลการทดลองดังตารางที่ 4.6

ตารางที่ 4.6 แสดงค่าความถูกต้องของเทคนิค Information Gain

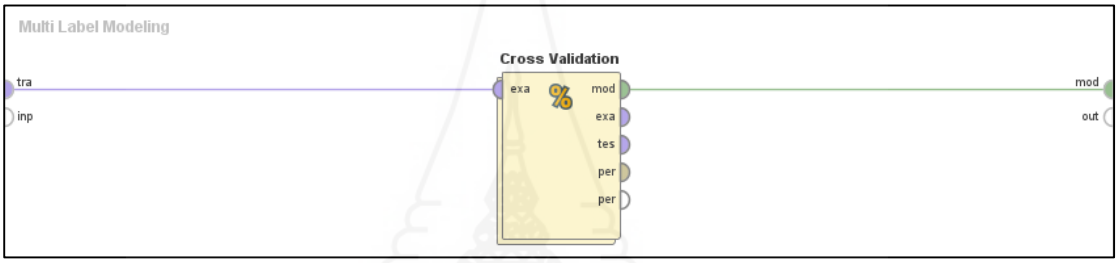
Target	Accuracy
A	0.859
B	0.940
C	0.974
Average	0.925

2.2.3 เทคนิค Gain Ratio

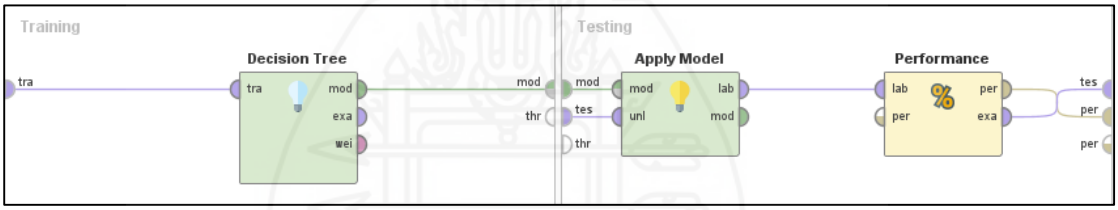
ผู้วิจัยทำการออกแบบการคัดเลือกคุณลักษณะสำคัญของเทคนิค Gain Ratio ใน selection3 ดังภาพที่ 4.12



(ก)



(ข)



(ค)

ภาพที่ 4.12 แสดงการออกแบบการคัดเลือกคุณลักษณะสำคัญของเทคนิค Gain Ratio

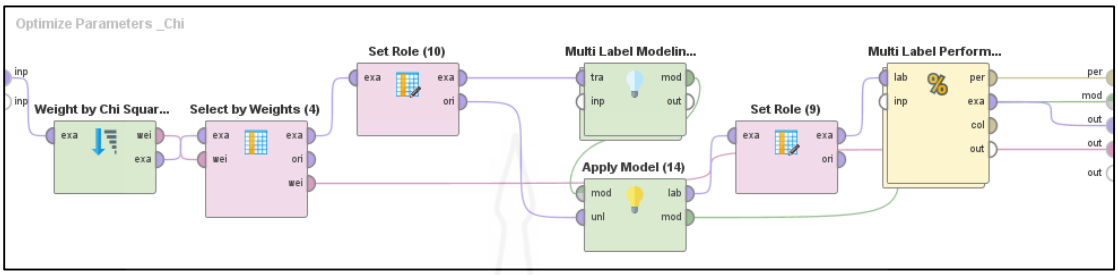
จากภาพที่ 4.12 แสดงการออกแบบการคัดเลือกคุณลักษณะสำคัญของเทคนิค Gain Ratio ของการจำแนกประเภทข้อมูลด้วยต้นไม้ตัดสินใจบนพื้นฐานเงื่อนไขหลายเป้าหมาย เพื่อทำการหาค่าความถูกต้องของเทคนิคการคัดเลือกคุณลักษณะสำคัญ สำหรับคัดเลือกเทคนิคการเลือกคุณลักษณะสำคัญที่เหมาะสมบนพื้นฐานเงื่อนไขหลายเป้าหมาย ได้ผลการทดลองดังตารางที่ 4.7

ตารางที่ 4.7 แสดงค่าความถูกต้องของเทคนิค Gain Ratio

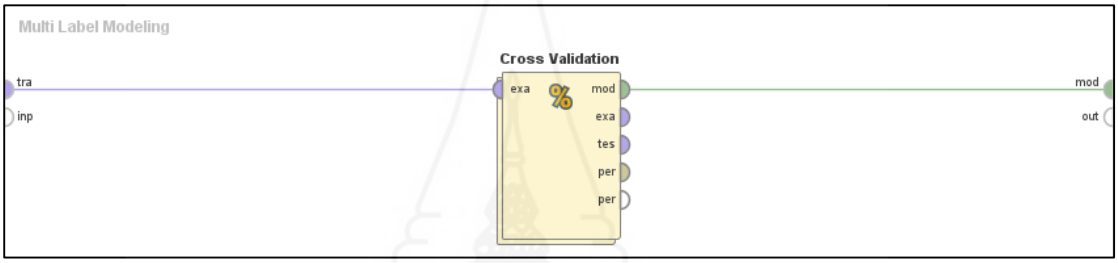
Target	Accuracy
A	0.833
B	0.897
C	0.957
Average	0.896

2.2.4 เทคนิค Chi Square

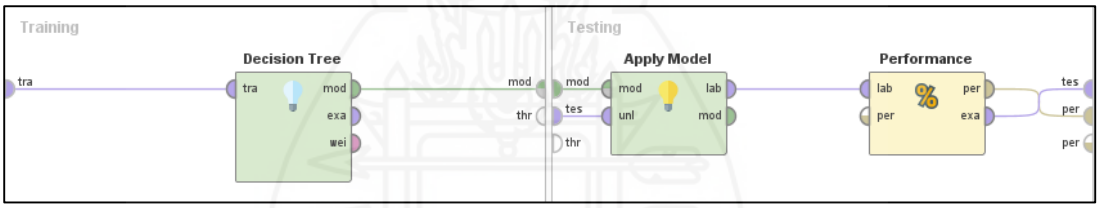
ผู้วิจัยทำการออกแบบการคัดเลือกคุณลักษณะสำคัญของเทคนิค Chi Square ใน selection4 ดังภาพที่ 4.13



(ก)



(ข)



(ค)

ภาพที่ 4.13 แสดงการออกแบบการคัดเลือกคุณลักษณะสำคัญของเทคนิค Chi Square

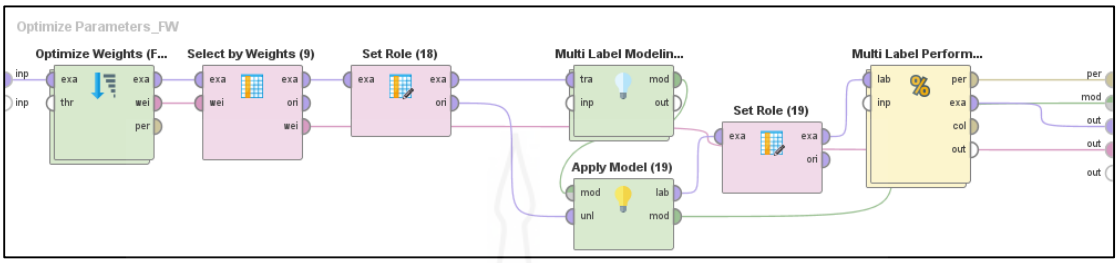
จากภาพที่ 4.13 แสดงการออกแบบการคัดเลือกคุณลักษณะสำคัญของเทคนิค Chi Square ของการจำแนกประเภทข้อมูลด้วยต้นไม้ตัดสินใจบนพื้นฐานเงื่อนไขหลายเป้าหมาย เพื่อทำการหาค่าความถูกต้องของเทคนิคการคัดเลือกคุณลักษณะสำคัญ สำหรับคัดเลือกเทคนิคการเลือกคุณลักษณะสำคัญที่เหมาะสมบนพื้นฐานเงื่อนไขหลายเป้าหมาย ได้ผลการทดลองดังตารางที่ 4.8

ตารางที่ 4.8 แสดงค่าความถูกต้องของเทคนิค Chi Square

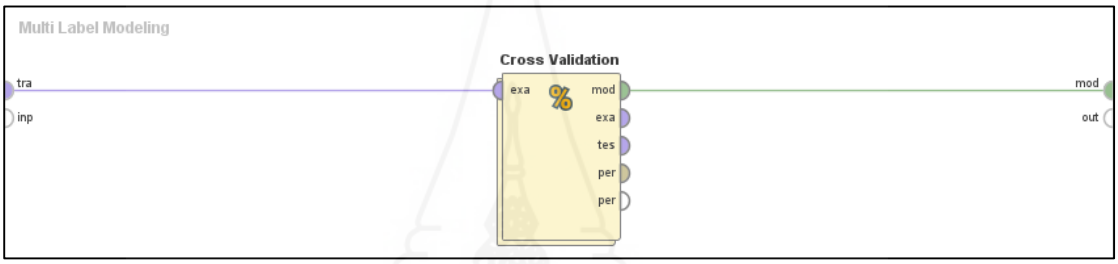
Target	Accuracy
A	0.867
B	0.940
C	0.974
Average	0.927

2.2.5 เทคนิค Forward Selection

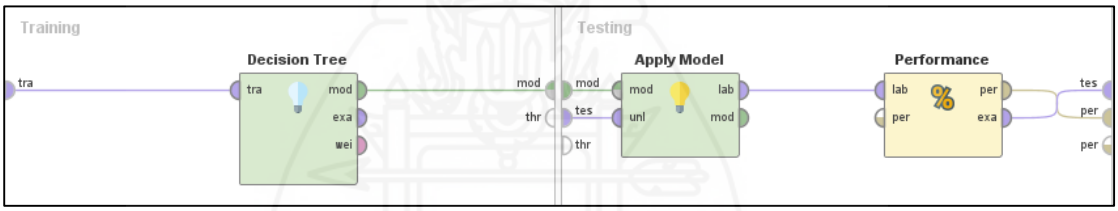
ผู้วิจัยทำการออกแบบการคัดเลือกคุณลักษณะสำคัญของเทคนิค Forward Selection ใน selection5 ดังภาพที่ 4.14



(ก)



(ข)



(ค)

ภาพที่ 4.14 แสดงการออกแบบการคัดเลือกคุณลักษณะสำคัญของเทคนิค Forward Selection

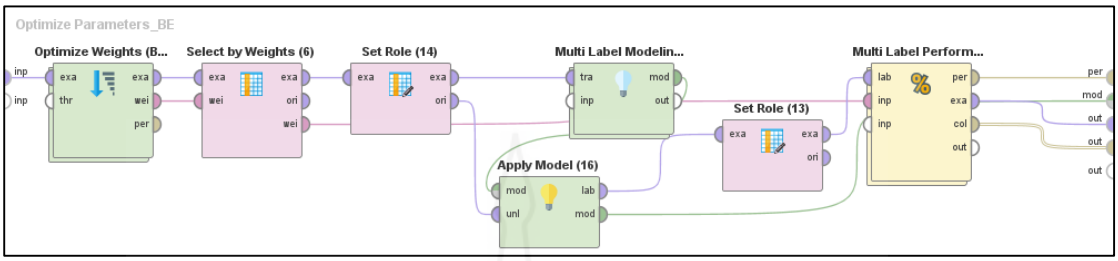
จากภาพที่ 4.14 แสดงการออกแบบการคัดเลือกคุณลักษณะสำคัญของเทคนิค Forward Selection ของการจำแนกประเภทข้อมูลด้วยต้นไม้ตัดสินใจบนพื้นฐานเงื่อนไขหลายเป้าหมาย เพื่อทำการหาค่าความถูกต้องของเทคนิคการคัดเลือกคุณลักษณะสำคัญ สำหรับคัดเลือกเทคนิคการเลือกคุณลักษณะสำคัญที่เหมาะสมบนพื้นฐานเงื่อนไขหลายเป้าหมาย ได้ผลการทดลองดังตารางที่ 4.9

ตารางที่ 4.9 แสดงค่าความถูกต้องของเทคนิค Forward Selection

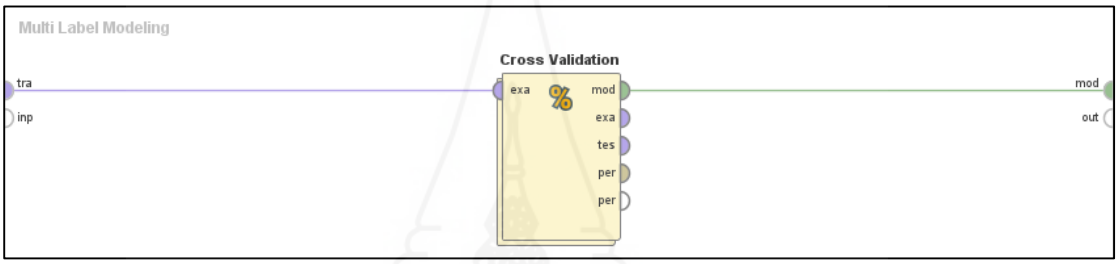
Target	Accuracy
A	0.847
B	0.924
C	0.989
Average	0.920

2.2.6 เทคนิค Backward Elimination

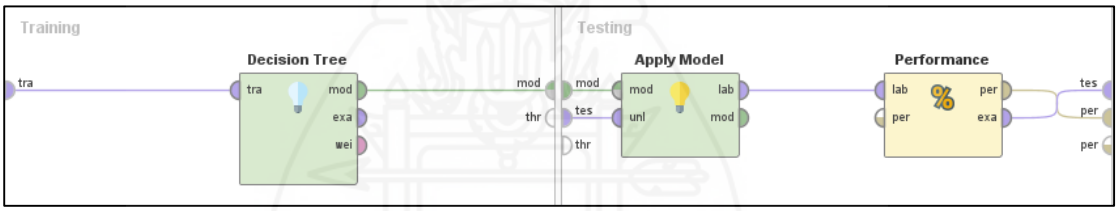
ผู้วิจัยทำการออกแบบการคัดเลือกคุณลักษณะสำคัญของเทคนิค Backward Elimination ใน selection6 ดังภาพที่ 4.15



(ก)



(ข)



(ค)

ภาพที่ 4.15 แสดงการออกแบบการคัดเลือกคุณลักษณะสำคัญของเทคนิค Backward Elimination

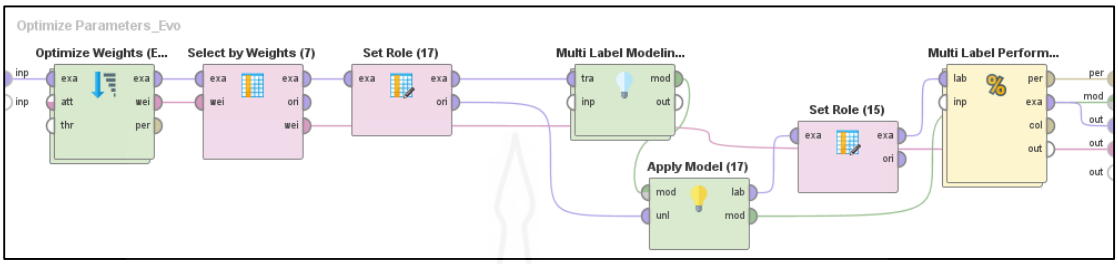
จากภาพที่ 4.15 แสดงการออกแบบการคัดเลือกคุณลักษณะสำคัญของเทคนิค Backward Elimination ของการจำแนกประเภทข้อมูลด้วยต้นไม้ตัดสินใจบนพื้นฐานเงื่อนไขหลายเป้าหมาย เพื่อทำการหาค่าความถูกต้องของเทคนิคการคัดเลือกคุณลักษณะสำคัญ สำหรับคัดเลือกเทคนิคการเลือกคุณลักษณะสำคัญที่เหมาะสมบนพื้นฐานเงื่อนไขหลายเป้าหมาย ได้ผลการทดลองดังตารางที่ 4.10

ตารางที่ 4.10 แสดงค่าความถูกต้องของเทคนิค Backward Elimination

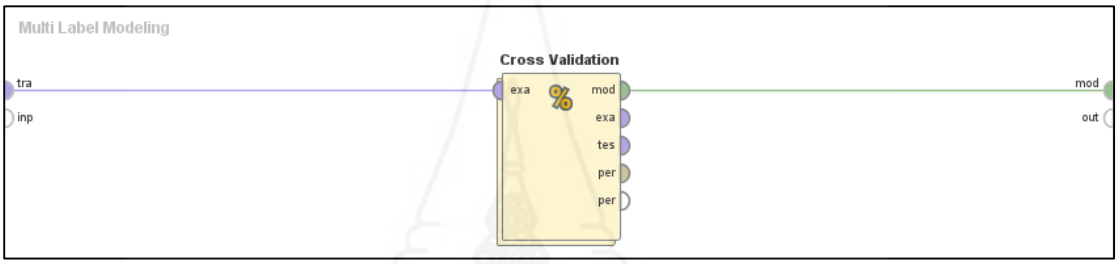
Target	Accuracy
A	0.913
B	0.946
C	0.962
Average	0.940

2.2.7 เทคนิค Evolutionary Selection

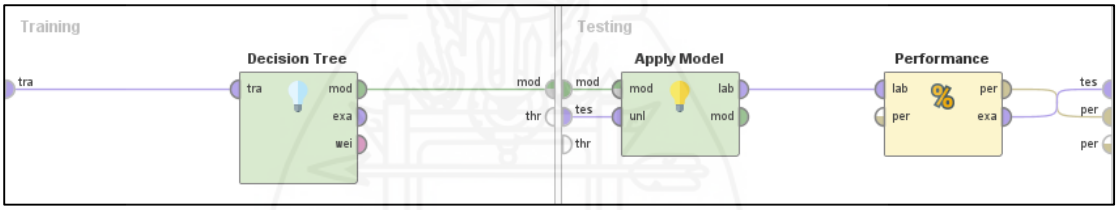
ผู้วิจัยทำการออกแบบการคัดเลือกคุณลักษณะสำคัญของเทคนิค Evolutionary Selection ใน selection7 ดังภาพที่ 4.16



(ก)



(ข)



(ค)

ภาพที่ 4.16 แสดงการออกแบบการคัดเลือกคุณลักษณะสำคัญของเทคนิค Evolutionary Selection

จากภาพที่ 4.16 แสดงการออกแบบการคัดเลือกคุณลักษณะสำคัญของเทคนิค Evolutionary Selection ของการจำแนกประเภทข้อมูลด้วยต้นไม้ตัดสินใจบนพื้นฐานเงื่อนไขหลายเป้าหมาย เพื่อทำการหาค่าความถูกต้องของเทคนิคการคัดเลือกคุณลักษณะสำคัญ สำหรับคัดเลือกเทคนิคการเลือกคุณลักษณะสำคัญที่เหมาะสมบนพื้นฐานเงื่อนไขหลายเป้าหมาย ได้ผลการทดลองดังตารางที่ 4.11

ตารางที่ 4.11 แสดงค่าความถูกต้องของเทคนิค Evolutionary Selection

Target	Accuracy
A	0.901
B	0.947
C	0.985
Average	0.944

จากนั้นผู้วิจัยทำการคัดเลือกเทคนิคการเลือกคุณลักษณะสำคัญที่เหมาะสมบนพื้นฐานเงื่อนไขหลายเป้าหมาย โดยทำการคัดเลือกด้วยการแปรผันไปตามเป้าหมาย จำนวน 3 เป้าหมาย ได้แก่ 1) เป้าหมาย A 2) เป้าหมาย B และ 3) เป้าหมาย C โดยใช้เทคนิคการคัดเลือกคุณลักษณะที่หลากหลาย ทั้งหมด 7 เทคนิค จากนั้นผู้วิจัยพิจารณาเลือกเทคนิคที่เหมาะสมจากค่าความถูกต้อง (accuracy) ของเป้าหมาย จำนวน 3 เป้าหมาย ได้แก่ 1) เป้าหมาย a 2) เป้าหมาย b และ 3) เป้าหมาย c แสดงดังตารางที่ 4.12

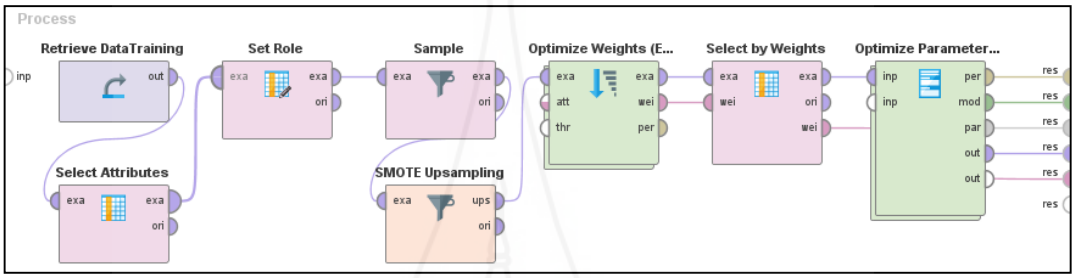
ตารางที่ 4.12 แสดงค่าความถูกต้องของเทคนิคการเลือกคุณลักษณะสำคัญ

Feature Selection Methods	Accuracy			Average Accuracy
	A	B	C	
Filter methods				
Correlation base Feature Selection	0.863	0.913	0.993	0.923
Information Gain	0.859	0.940	0.974	0.925
Gain Ratio	0.833	0.897	0.957	0.896
Chi Square	0.867	0.940	0.974	0.927
Wrapper approach				
Forward Selection	0.847	0.924	0.989	0.920
Backward Elimination	0.913	0.946	0.962	0.940
Evolutionary Selection	**0.901**	**0.947**	**0.985**	**0.944**

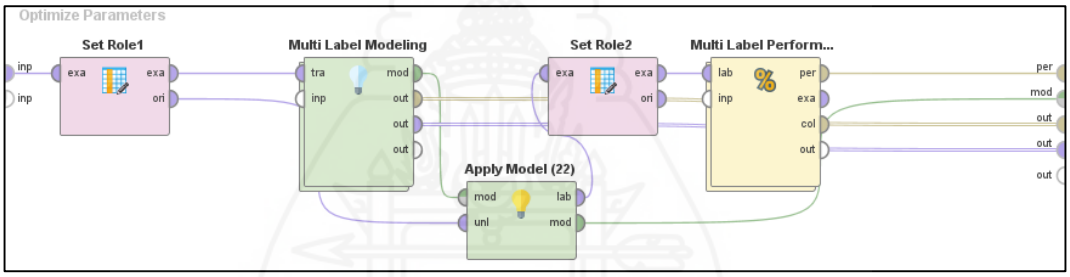
จากตารางที่ 4.12 พบว่า เทคนิค Evolutionary Selection ให้ค่าความถูกต้องดีที่สุด โดยมีค่าความถูกต้องเฉลี่ยอยู่ที่ 0.944 ผู้วิจัยจึงเลือกใช้เทคนิค Evolutionary Selection สำหรับนำไปใช้ในการเลือกคุณลักษณะสำคัญแบบพลวัตกับการจำแนกต้นไม้ตัดสินใจบนพื้นฐานเงื่อนไขหลายเป้าหมาย

2.3 การเลือกคุณลักษณะสำคัญแบบพลวัต (dynamic feature selection)

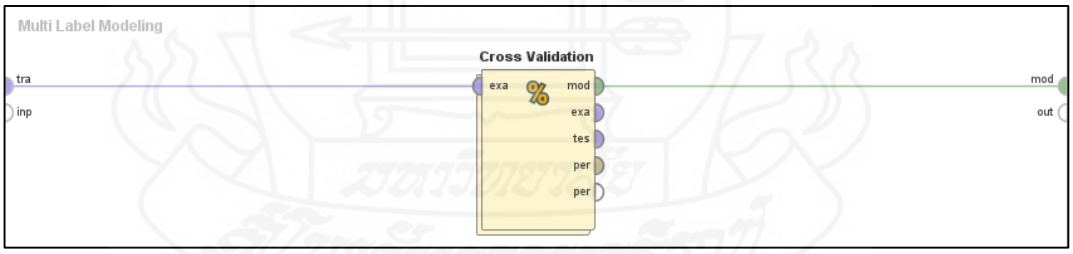
ผู้วิจัยจึงเลือกใช้เทคนิค Evolutionary Selection สำหรับนำไปใช้ในการเลือกคุณลักษณะสำคัญแบบพลวัตกับการจำแนกต้นไม้ตัดสินใจบนพื้นฐานเงื่อนไขหลายเป้าหมาย ทำการเลือกคุณลักษณะสำคัญแบบพลวัต ซึ่งคุณลักษณะสำคัญที่ได้จะแปรผันไปตามเป้าหมาย จำนวน 3 เป้าหมาย ได้แก่ 1) เป้าหมาย A แทนทุนได้เปล่า 2) เป้าหมาย B แทนทุนกู้ยืมเพื่อการศึกษา และ 3) เป้าหมาย C แทนทุนขาดแคลนทุนทรัพย์ โดยได้ออกแบบกระบวนการเลือกคุณลักษณะสำคัญแบบพลวัตบนพื้นฐานเงื่อนไขหายเป้าหมาย ภาพรวมกระบวนการดังภาพที่ 4.17



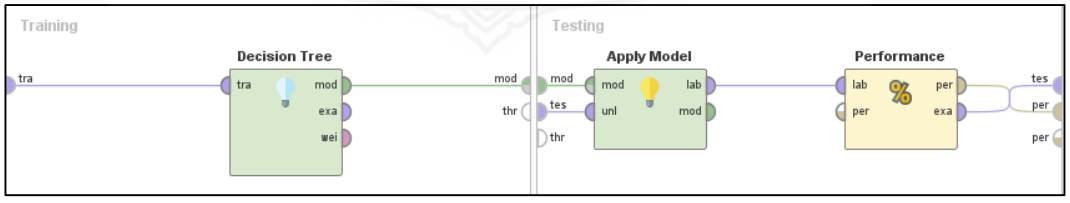
(ก)



(ข)



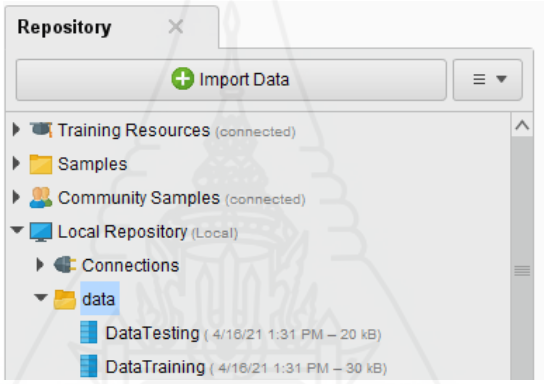
(ค)



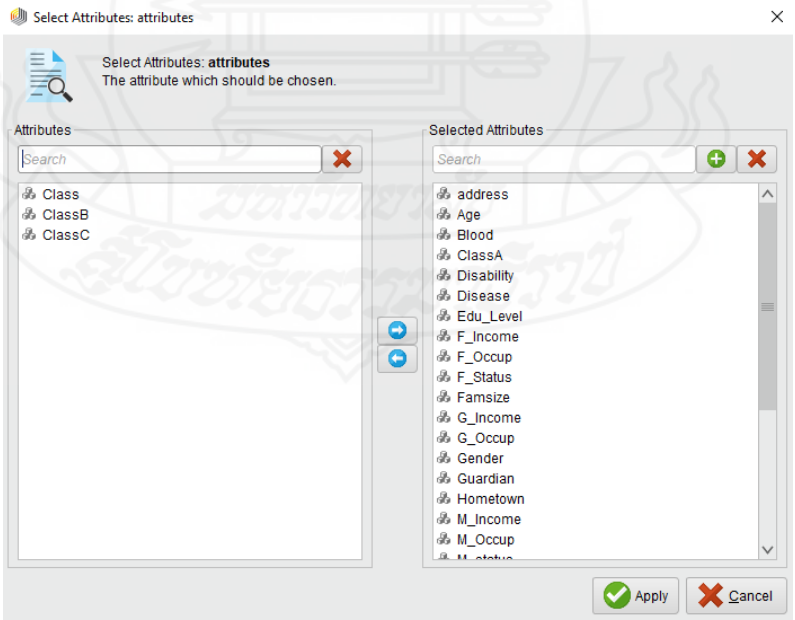
(ง)

ภาพที่ 4.17 แสดงภาพรวมกระบวนการเลือกคุณลักษณะสำคัญแบบพลวัตกับการจำแนกต้นไม้ตัดสินใจ

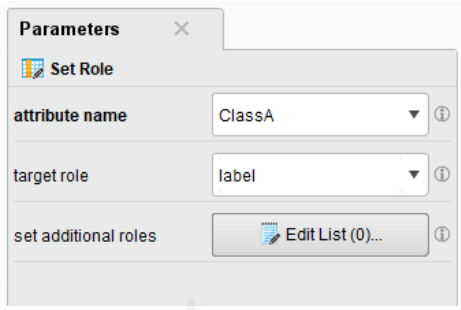
จากภาพที่ 4.17 แสดงกระบวนการนำเข้าข้อมูล (training data) โดยใช้ retrieve ดังภาพที่ 4.18 ทำการเลือกคุณลักษณะของข้อมูลโดยใช้ select attributes ดังภาพที่ 4.19 และกำหนดเป้าหมาย โดยใช้ set role เพื่อทำการปรับเปลี่ยนเป้าหมายที่ต้องการ ดังภาพที่ 4.20 จากนั้นทำการปรับสมดุลข้อมูลด้วยวิธี SMOTE ดังภาพที่ 4.21 -4.22 เลือกใช้เทคนิค Evolutionary Selection ในการคัดเลือกคุณลักษณะและหาค่าน้ำหนักของแต่ละคุณลักษณะออกมา จากนั้นทำการเลือกคุณลักษณะจากค่าน้ำหนักที่ได้สำหรับนำไปสร้างแบบจำลองโดยใช้ select by weights ดังภาพที่ 4.23 และกำหนด optimize parameters (grid) เพื่อใช้สำหรับกำหนดค่าพารามิเตอร์ที่เหมาะสมที่สุด ดังภาพที่ 4.24 และใช้การสร้างแบบจำลองแบบ multi label modeling ซึ่งภายใต้ multi label modeling กำหนด cross validation เป็นค่า 10-fold cross validation ดังภาพที่ 4.25 และสร้างแบบจำลองการจำแนกต้นไม้ตัดสินใจ



ภาพที่ 4.18 แสดงการนำเข้าข้อมูล (training data)

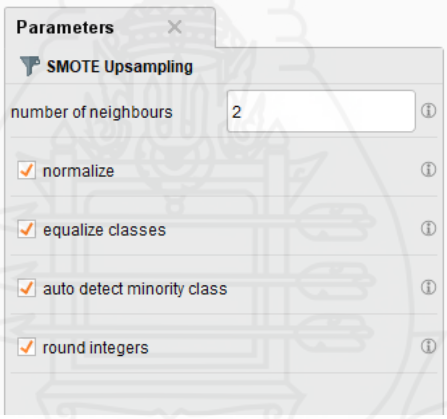


ภาพที่ 4.19 แสดงการเลือกคุณลักษณะของข้อมูล

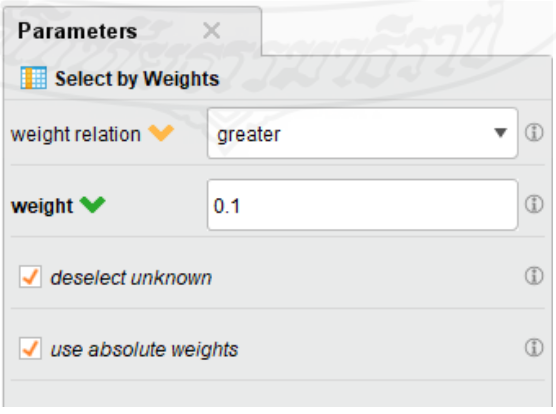


ภาพที่ 4.20 แสดงการกำหนดเป้าหมาย

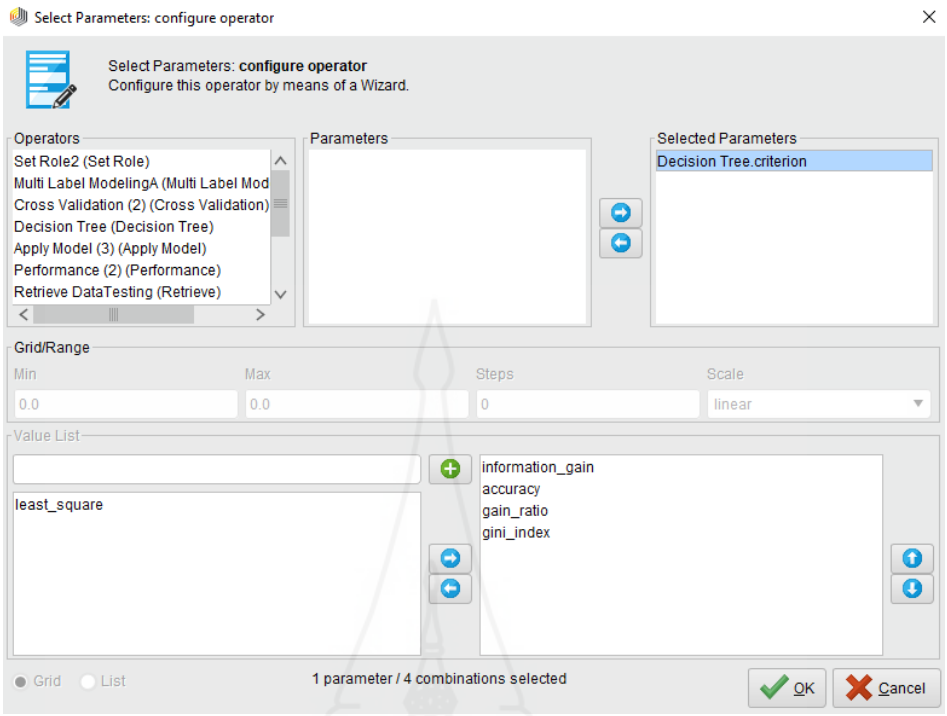
ภาพที่ 4.21 แสดงการกำหนดจำนวนข้อมูล sample



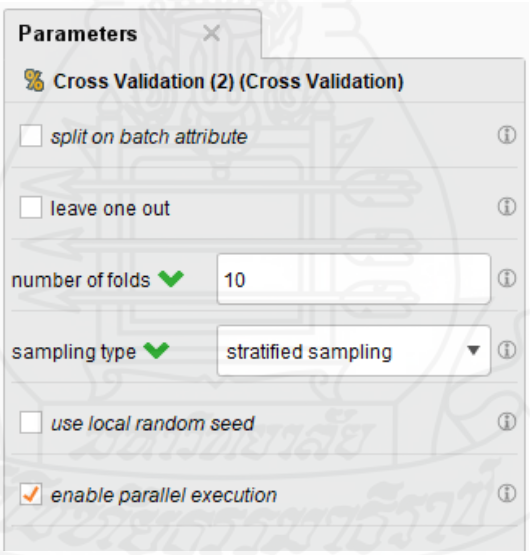
ภาพที่ 4.22 แสดงการปรับสมดุลข้อมูลด้วยวิธี SMOTE



ภาพที่ 4.23 แสดงการกำหนดค่าพารามิเตอร์ select by weights



ภาพที่ 4.24 แสดงการกำหนด optimize parameters (grid)



ภาพที่ 4.25 แสดงการกำหนด cross validation เท่ากับ 10

จากนั้นทำการเลือกคุณลักษณะสำคัญแบบพลวัต โดยกำหนด set role เป็นเป้าหมาย A และทำการเปลี่ยนเป้าหมายเป็นเป้าหมาย B และเป้าหมาย C ตามลำดับ เพื่อทำการปรับเปลี่ยนเป้าหมายที่ต้องการคัดเลือกคุณลักษณะสำคัญแบบพลวัตที่แปรผันไปตามเป้าหมายที่ต้องการ ได้ผลการทดลองดังตารางที่ 4.13 - 4.15

ตารางที่ 4.13 แสดงค่าน้ำหนักของคุณลักษณะของเป้าหมาย A

No	Attribute	Weight
1.	M_Income	1.000
2.	Nationality	0.629
3.	F_Status	0.627
4.	Major	0.565
5.	Talent	0.528
6.	Guardian	0.512
7.	Religion	0.491
8.	Disability	0.481
9.	Son_number	0.454
10.	P_status	0.408
11.	Recruit	0.391
12.	Race	0.380
13.	Famsize	0.369
14.	Age	0.328
15.	Disease	0.301

No	Attribute	Weight
16.	F_Income	0.300
17.	Old_Gpax	0.284
18.	M_status	0.268
19.	Hometown	0.264
20.	Quota	0.263
21.	F_Occup	0.247
22.	address	0.212
23.	M_Occup	0.197
24.	Gender	0.158
25.	Blood	0.140
26.	G_Income	0.118
27.	Old_Edu	0.101
28.	G_Occup	0.028
29.	Edu_Level	0.000

จากตารางที่ 4.13 แสดงค่าน้ำหนักของคุณลักษณะของเป้าหมาย A ที่ได้จากการคัดเลือกคุณลักษณะสำคัญด้วยเทคนิค Evolutionary Selection โดยมีค่าน้ำหนักของคุณลักษณะต่ำสุดอยู่ที่ 0.000 และค่าน้ำหนักสูงสุดอยู่ที่ 1.000 ผู้วิจัยกำหนดให้คัดเลือกคุณลักษณะสำคัญจากค่าน้ำหนักโดยใช้ select by weights ซึ่งการกำหนดค่าพารามิเตอร์ เท่ากับ 0.1 ในการเลือกคุณลักษณะสำคัญสำหรับนำไปสร้างแบบจำลองจำแนกต้นไม้ตัดสินใจ

ตารางที่ 4.14 แสดงค่าน้ำหนักของคุณลักษณะของเป้าหมาย B

No	Attribute	Weight
1.	F_Status	1.000
2.	Guardian	0.798
3.	F_Occup	0.545
4.	Hometown	0.385
5.	Nationality	0.380
6.	Disease	0.339
7.	Talent	0.332
8.	Major	0.305
9.	address	0.302
10.	Gender	0.288
11.	Age	0.281
12.	Son_number	0.265
13.	G_Occup	0.235
14.	F_Income	0.215
15.	M_Occup	0.213

No	Attribute	Weight
16.	Race	0.211
17.	Disability	0.186
18.	Old_Gpax	0.169
19.	P_status	0.116
20.	G_Income	0.113
21.	M_Income	0.096
22.	M_status	0.093
23.	Recruit	0.086
24.	Blood	0.082
25.	Quota	0.058
26.	Religion	0.042
27.	Edu_Level	0.023
28.	Old_Edu	0.012
29.	Famsize	0.000

จากตารางที่ 4.14 แสดงค่าน้ำหนักของคุณลักษณะของเป้าหมาย B ที่ได้จากการคัดเลือกคุณลักษณะสำคัญด้วยเทคนิค Evolutionary Selection โดยมีค่าน้ำหนักของคุณลักษณะต่ำสุดอยู่ที่ 0.000 และค่าน้ำหนักสูงสุดอยู่ที่ 1.000 ผู้วิจัยกำหนดให้คัดเลือกคุณลักษณะสำคัญจากค่าน้ำหนักโดยใช้ select by weights ซึ่งการกำหนดค่าพารามิเตอร์ เท่ากับ 0.1 ในการเลือกคุณลักษณะสำคัญสำหรับนำไปสร้างแบบจำลองจำแนกต้นไม้ตัดสินใจ

ตารางที่ 4.15 แสดงค่าน้ำหนักของคุณลักษณะของเป้าหมาย C

No	Attribute	Weight
1.	Blood	1.000
2.	Religion	0.903
3.	Hometown	0.881
4.	G_Income	0.799
5.	Major	0.673
6.	F_Status	0.668
7.	Quota	0.663
8.	P_status	0.518
9.	M_Income	0.509
10.	Recruit	0.424
11.	F_Income	0.373
12.	Age	0.313
13.	Gender	0.287
14.	M_Occup	0.285
15.	Old_Gpax	0.277

No	Attribute	Weight
16.	Nationality	0.255
17.	G_Occup	0.242
18.	Disease	0.186
19.	Race	0.177
20.	Disability	0.134
21.	address	0.099
22.	Old_Edu	0.076
23.	Guardian	0.076
24.	Famsize	0.062
25.	M_status	0.019
26.	Talent	0.007
27.	F_Occup	0.004
28.	Son_number	0.001
29.	Edu_Level	0.000

จากตารางที่ 4.15 แสดงค่าน้ำหนักของคุณลักษณะของเป้าหมาย C ที่ได้จากการคัดเลือกคุณลักษณะสำคัญด้วยเทคนิค Evolutionary Selection โดยมีค่าน้ำหนักของคุณลักษณะต่ำสุดอยู่ที่ 0.000 และค่าน้ำหนักสูงสุดอยู่ที่ 1.000 ผู้วิจัยกำหนดให้คัดเลือกคุณลักษณะสำคัญจากค่าน้ำหนักโดยใช้ select by weights ซึ่งการกำหนดค่าพารามิเตอร์ เท่ากับ 0.1 ในการเลือกคุณลักษณะสำคัญสำหรับนำไปสร้างแบบจำลองจำแนกต้นไม้ตัดสินใจ

จากตารางที่ 4.13 – 4.15 แสดงค่าน้ำหนักของคุณลักษณะของแต่ละเป้าหมาย ผู้วิจัยทำการเลือกคุณลักษณะจากค่าน้ำหนักที่ได้สำหรับนำไปสร้างแบบจำลองโดยใช้ select by weights โดยการกำหนดค่าพารามิเตอร์ เท่ากับ 0.1 ในการเลือกคุณลักษณะสำคัญจาก 29 คุณลักษณะของเป้าหมาย จำนวน 3 เป้าหมาย ซึ่งได้ผลการทดลองดังตารางที่ 4.16 - 4.17

ตารางที่ 4.16 แสดงค่าน้ำหนักของแต่ละคุณลักษณะที่ได้จากการเลือกคุณลักษณะสำคัญแบบพลวัต

Target A		Target B		Target C	
Attribute	Weight	Attribute	Weight	Attribute	Weight
M_Income	**1.000**	**F_Status**	**1.000**	**Blood**	**1.000**
Nationality	0.629	Guardian	0.798	Religion	0.903
F_Status	0.627	F_Occup	0.545	Hometown	0.881
Major	0.565	Hometown	0.385	G_Income	0.799
Talent	0.528	Nationality	0.380	Major	0.673
Guardian	0.512	Disease	0.339	F_Status	0.668
Religion	0.491	Talent	0.332	Quota	0.663
Disability	0.481	Major	0.305	P_status	0.518
Son_number	0.454	address	0.302	M_Income	0.509
P_status	0.408	Gender	0.288	Recruit	0.424
Recruit	0.391	Age	0.281	F_Income	0.373
Race	0.380	Son_number	0.265	Age	0.313
Famsize	0.369	G_Occup	0.235	Gender	0.287
Age	0.328	F_Income	0.215	M_Occup	0.285
Disease	0.301	M_Occup	0.213	Old_Gpax	0.277
F_Income	0.300	Race	0.211	Nationality	0.255
Old_Gpax	0.284	Disability	0.186	G_Occup	0.242
M_status	0.268	Old_Gpax	0.169	Disease	0.186
Hometown	0.264	P_status	0.116	Race	0.177
Quota	0.263	**G_Income**	**0.113**	**Disability**	**0.134**
F_Occup	0.247				
address	0.212				
M_Occup	0.197				
Gender	0.158				
Blood	0.140				
G_Income	0.118				
Old_Edu	**0.101**				

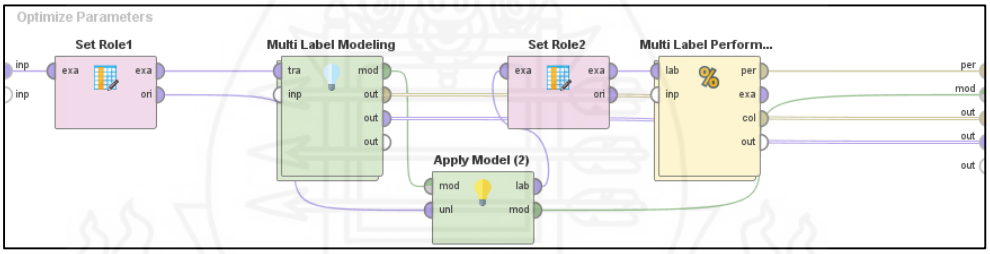
ตารางที่ 4.17 แสดงผลการเลือกคุณลักษณะสำคัญแบบพลวัต

target	attribute selection	accuracy
A	27	0.918
B	20	0.939
C	20	0.959

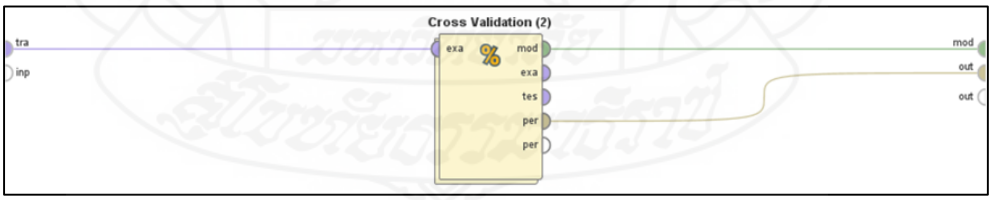
จากตารางที่ 4.17 ผลการคัดเลือกคุณลักษณะสำคัญแบบพลวัต พบว่า สามารถคัดเลือกคุณลักษณะสำคัญของแต่ละเป้าหมายได้แตกต่างกัน โดยสามารถลดจำนวนคุณลักษณะสำคัญจาก 29 คุณลักษณะ ของเป้าหมาย A เหลือจำนวน 27 คุณลักษณะ ส่วนเป้าหมาย B เหลือจำนวน 20 คุณลักษณะ และเป้าหมาย C เหลือจำนวน 20 คุณลักษณะ

3. การสร้างแบบจำลอง (modeling)

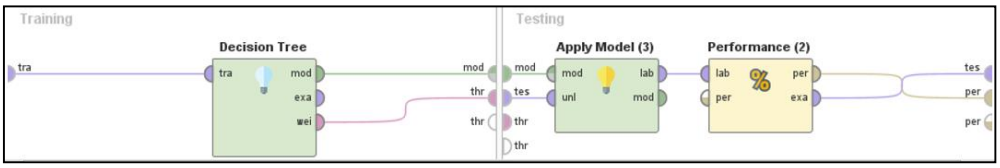
ผู้วิจัยนำชุดข้อมูล แบ่งเป็น 100, 300 และ 500 แถว เพื่อทดสอบแบบจำลอง โดยชุดข้อมูล 100, 300 และ 500 แบ่งเป็น 80:20 ทั้งหมด คือชุดข้อมูลฝึกสอนร้อยละ 80 และชุดข้อมูลทดสอบร้อยละ 20 และสร้างแบบจำลองการจำแนกประเภทข้อมูลด้วยเทคนิคต้นไม้ตัดสินใจ (decision tree) โดยใช้วิธี 10-fold cross validation เพื่อให้ข้อมูลทุกตัวมีโอกาสเป็นชุดข้อมูลฝึกสอน (training data) และชุดข้อมูลทดสอบ (testing data) ภาพรวมกระบวนการดังภาพที่ 4.26



(ก)



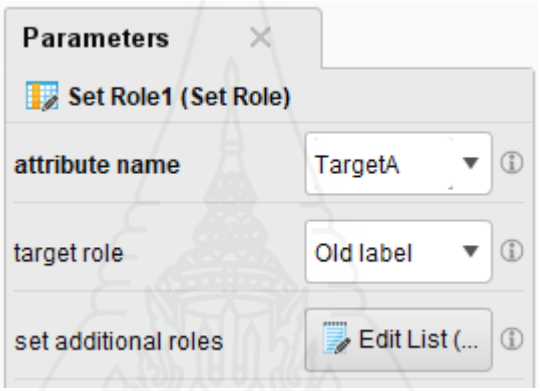
(ข)



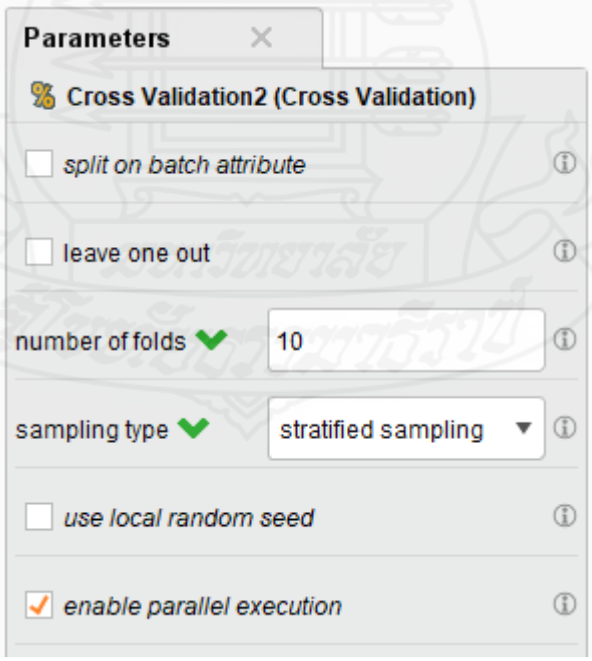
(ค)

ภาพที่ 4.26 แสดงภาพรวมกระบวนการสร้างแบบจำลองการจำแนกประเภทข้อมูลด้วยเทคนิคต้นไม้ตัดสินใจ (decision tree)

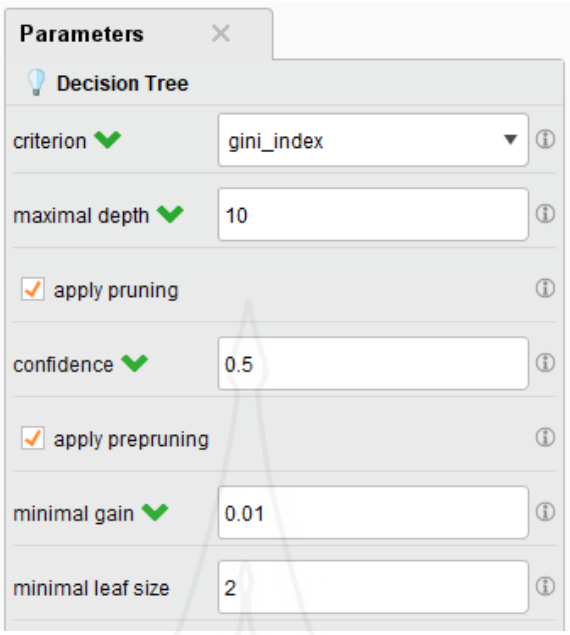
จากภาพที่ 4.26 แสดงกระบวนการสร้างแบบจำลองการจำแนกประเภทข้อมูลด้วยเทคนิคต้นไม้ตัดสินใจ (decision tree) โดยนำเข้าข้อมูลจากการเลือกคุณลักษณะ (ในภาพที่ 4.17) กำหนดเป้าหมายโดยใช้ set role1 ให้เป็นเป้าหมายเดียวกับการคัดเลือกคุณลักษณะ ดังภาพที่ 4.27 ใช้การสร้างแบบจำลองแบบ multi label modeling ซึ่งภายใต้ multi label modeling กำหนด cross validation เป็นค่า 10-fold cross validation ดังภาพที่ 4.28 เลือกใช้อัลกอริทึมต้นไม้ตัดสินใจ (decision tree) เพื่อเรียนรู้และทดสอบ ดังภาพที่ 4.29 และกำหนด performance ในการวัดประสิทธิภาพแบบจำลอง ในส่วนของการ testing เป็นการนำเข้าชุดข้อมูลทดสอบ 20% โดยใช้ retrieve กำหนด set role2 ในการกำหนดเป้าหมายการพยากรณ์ โดยใช้ apply model (2) สำหรับการทดสอบ และวัดค่าประสิทธิภาพด้วย multi label performance



ภาพที่ 4.27 แสดงการกำหนดเป้าหมายโดยใช้ set role2



ภาพที่ 4.28 แสดงการกำหนด cross validation เท่ากับ 10



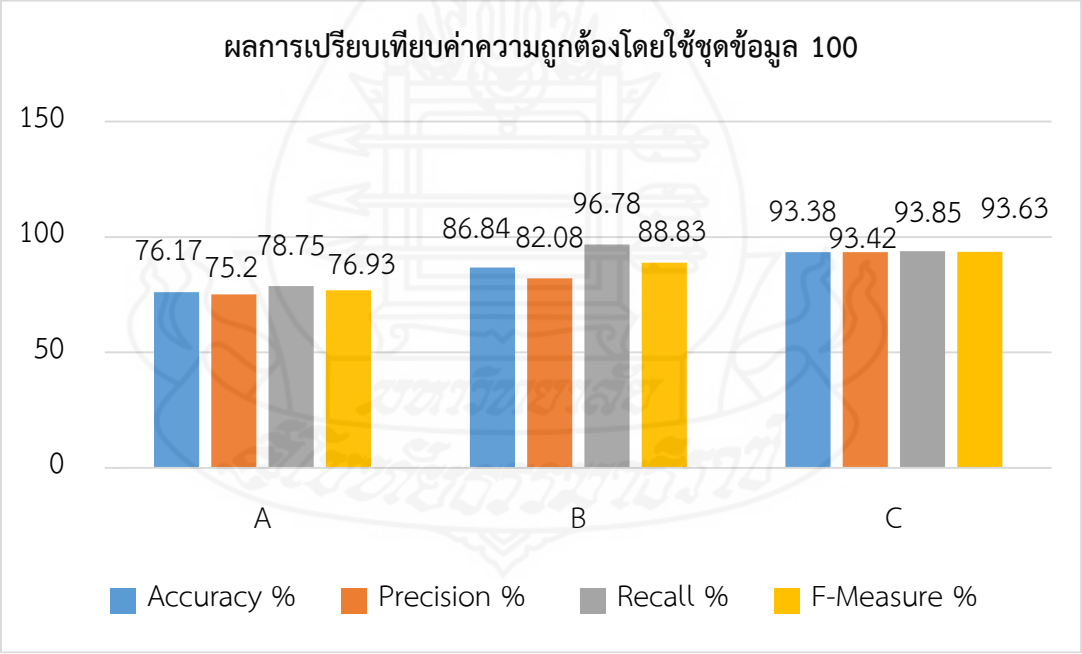
ภาพที่ 4.29 แสดงการสร้างแบบจำลองการจำแนกต้นไม้ตัดสินใจ

3.1 ผลการสร้างแบบจำลองโดยใช้ชุดข้อมูล จำนวน 100 แถว มาแบ่งออกเป็น 2 ส่วน คือ ชุดข้อมูลฝึกสอนร้อยละ 80 (จำนวน 80 แถว) และชุดข้อมูลทดสอบร้อยละ 20 (จำนวน 20 แถว) มาสร้างแบบจำลองการจำแนกประเภทข้อมูลด้วยเทคนิคต้นไม้ตัดสินใจ (decision tree) ได้ผลการทดลองดังตารางที่ 4.18

ตารางที่ 4.18 แสดงผลการวัดประสิทธิภาพของแบบจำลองด้วยชุดข้อมูล 100 แถว

target	accuracy %	precision %	recall %	f-measure %
A	76.17	75.20	78.75	76.93
B	86.84	82.08	96.78	88.83
C	93.38	93.42	93.85	93.63
average	**85.46**	**83.57**	**89.79**	**86.46**

จากตารางที่ 4.18 ผลการวัดประสิทธิภาพแบบจำลองด้วยชุดข้อมูล 100 แถว พบว่า แบบจำลองการจำแนกต้นไม้ตัดสินใจบนพื้นฐานเงื่อนไขหลายเป้าหมาย ให้ค่าความถูกต้องร้อยละ 85.46 ค่าความแม่นยำร้อยละ 83.57 ค่าเรียกคืนร้อยละ 89.79 และการวัดประสิทธิภาพโดยรวมร้อยละ 86.46



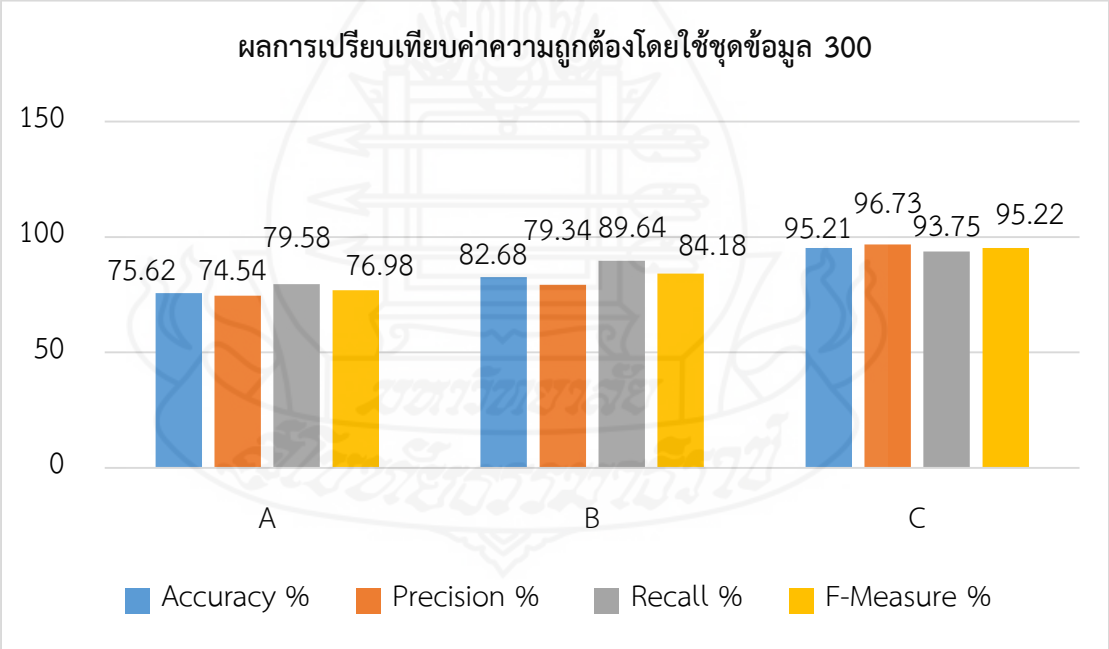
ภาพที่ 4.30 แสดงผลการเปรียบเทียบค่าความถูกต้องโดยใช้ชุดข้อมูล 100

3.2 ผลการสร้างแบบจำลองโดยใช้ชุดข้อมูล จำนวน 300 แถว มาแบ่งออกเป็น 2 ส่วน คือ ชุดข้อมูลฝึกสอนร้อยละ 80 (จำนวน 240 แถว) และชุดข้อมูลทดสอบร้อยละ 20 (จำนวน 60 แถว) มาสร้างแบบจำลองการจำแนกประเภทข้อมูลด้วยเทคนิคต้นไม้ตัดสินใจ (decision tree) ได้ผลการทดลองดังตารางที่ 4.19

ตารางที่ 4.19 แสดงผลการวัดประสิทธิภาพของแบบจำลองด้วยชุดข้อมูล 300 แถว

target	accuracy %	precision %	recall %	f-measure %
A	75.62	74.54	79.58	76.98
B	82.68	79.34	89.64	84.18
C	95.21	96.73	93.75	95.22
average	**84.50**	**83.54**	**87.66**	**85.46**

จากตารางที่ 4.19 ผลการวัดประสิทธิภาพแบบจำลองด้วยชุดข้อมูล 300 แถว พบว่า แบบจำลองการจำแนกต้นไม้ตัดสินใจบนพื้นฐานเงื่อนไขหลายเป้าหมาย ให้ค่าความถูกต้องร้อยละ 84.50 ค่าความแม่นยำร้อยละ 83.54 ค่าเรียกคืนร้อยละ 87.66 และการวัดประสิทธิภาพโดยรวมร้อยละ 85.46



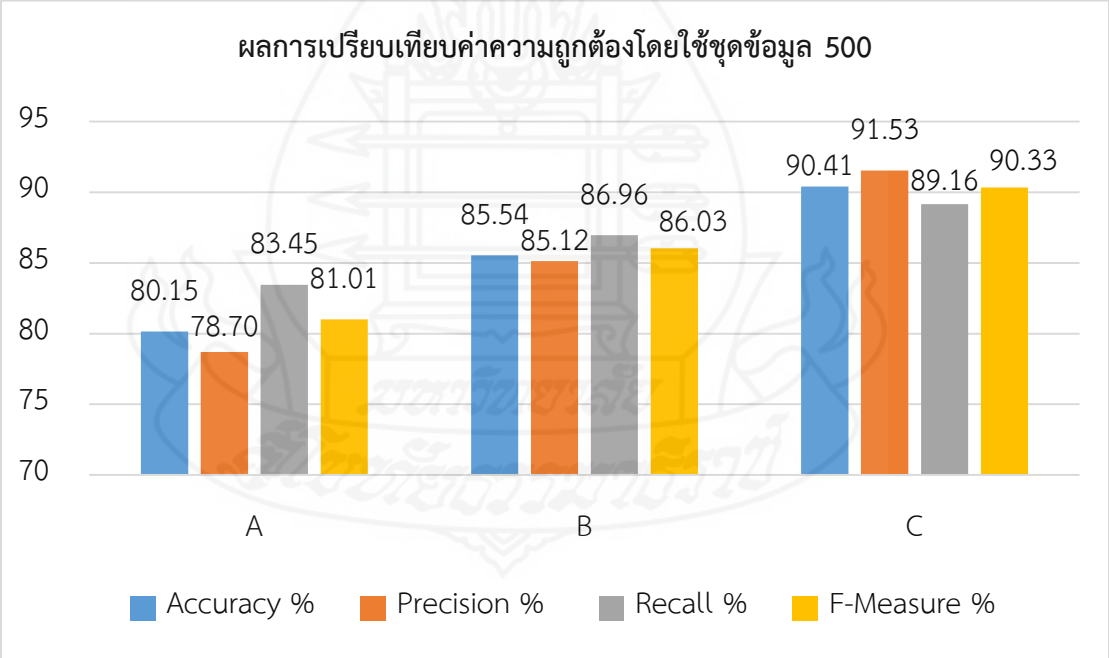
ภาพที่ 4.31 แสดงผลการเปรียบเทียบค่าความถูกต้องโดยใช้ชุดข้อมูล 300

3.3 ผลการสร้างแบบจำลองโดยใช้ชุดข้อมูล จำนวน 500 แถว มาแบ่งออกเป็น 2 ส่วน คือ ชุดข้อมูลฝึกสอนร้อยละ 80 (จำนวน 400 แถว) และชุดข้อมูลทดสอบร้อยละ 20 (จำนวน 100 แถว) มาสร้างแบบจำลองการจำแนกประเภทข้อมูลด้วยเทคนิคต้นไม้ตัดสินใจ (decision tree) ได้ผลการทดลองดังตารางที่ 4.20

ตารางที่ 4.20 แสดงผลการวัดประสิทธิภาพของแบบจำลองด้วยชุดข้อมูล 500 แถว

target	accuracy %	precision %	recall %	f-measure %
A	80.15	78.70	83.45	81.01
B	85.54	85.12	86.96	86.03
C	**90.41**	**91.53**	**89.16**	**90.33**
average	**85.37**	**85.12**	**86.52**	**85.79**

จากตารางที่ 4.20 ผลการวัดประสิทธิภาพแบบจำลองด้วยชุดข้อมูล 500 แถว พบว่า แบบจำลองการจำแนกต้นไม้ตัดสินใจบนพื้นฐานเงื่อนไขหลายเป้าหมาย ให้ค่าความถูกต้องร้อยละ 85.37 ค่าความแม่นยำร้อยละ 85.12 ค่าเรียกคืนร้อยละ 86.52 และการวัดประสิทธิภาพโดยรวมร้อยละ 85.79

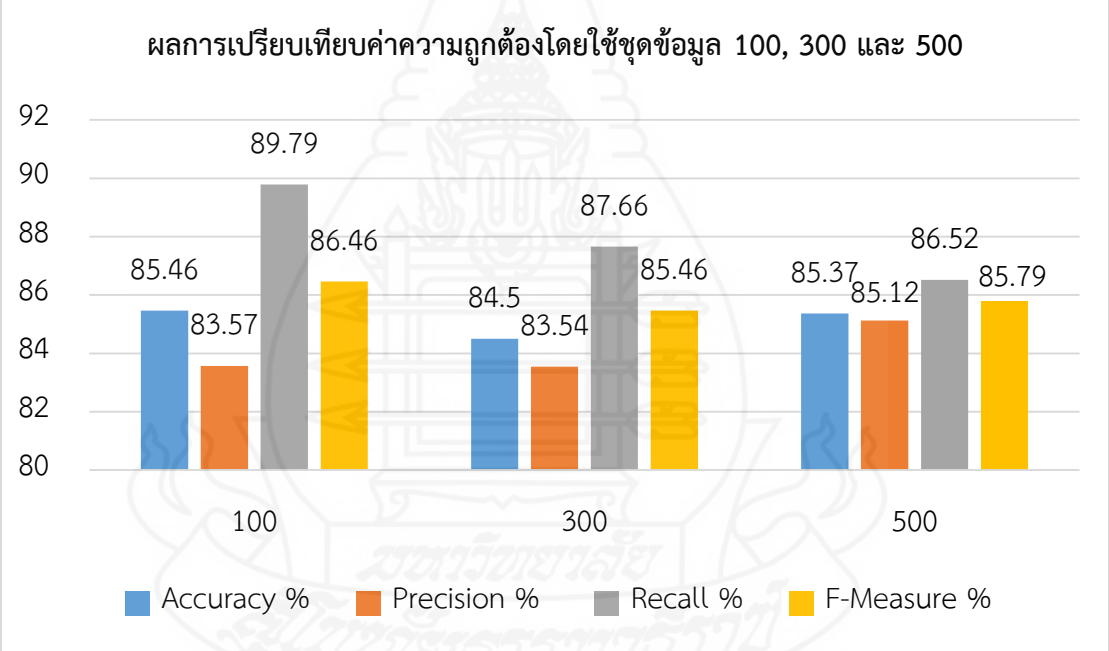


ภาพที่ 4.32 แสดงผลการเปรียบเทียบค่าความถูกต้องโดยใช้ชุดข้อมูล 500

3.4 ผลการเปรียบเทียบการนำชุดข้อมูล มาแบ่งเป็น 100, 300 และ 500 แถว เพื่อทดสอบแบบจำลอง โดยชุดข้อมูล 100, 300 และ 500 ได้ผลการเปรียบเทียบค่าความถูกต้องดังตารางที่ 4.21 ตารางที่ 4.21 แสดงผลการเปรียบเทียบค่าความถูกต้องของแบบจำลอง

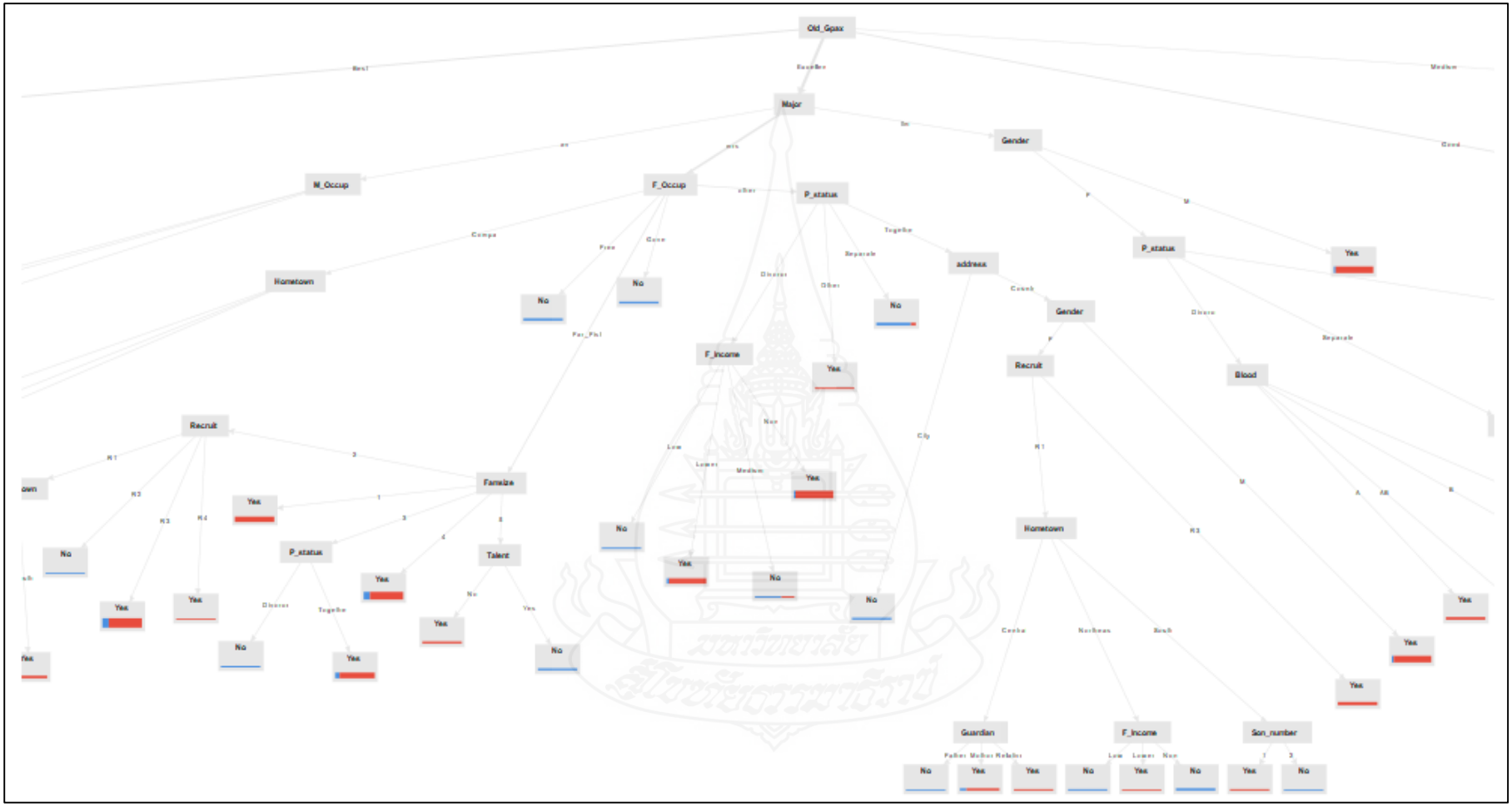
dataset	accuracy %	precision %	recall %	f-measure %
100	85.46	83.57	89.79	86.46
300	84.50	83.54	87.66	85.46
500	85.37	85.12	86.52	85.79

จากตารางที่ 4.21 ผลการวัดประสิทธิภาพแบบจำลองด้วยชุดข้อมูล 100, 300 และ 500 แถว พบว่า ชุดข้อมูลจำนวน 500 แถว เมื่อนำมาสร้างแบบจำลองการจำแนกต้นไม้ตัดสินใจ ให้ค่าความถูกต้องของแบบจำลองร้อยละ 85.37 ค่าความแม่นยำร้อยละ 85.12 ค่าเรียกคืนร้อยละ 86.52 และการวัดประสิทธิภาพโดยรวมร้อยละ 85.79

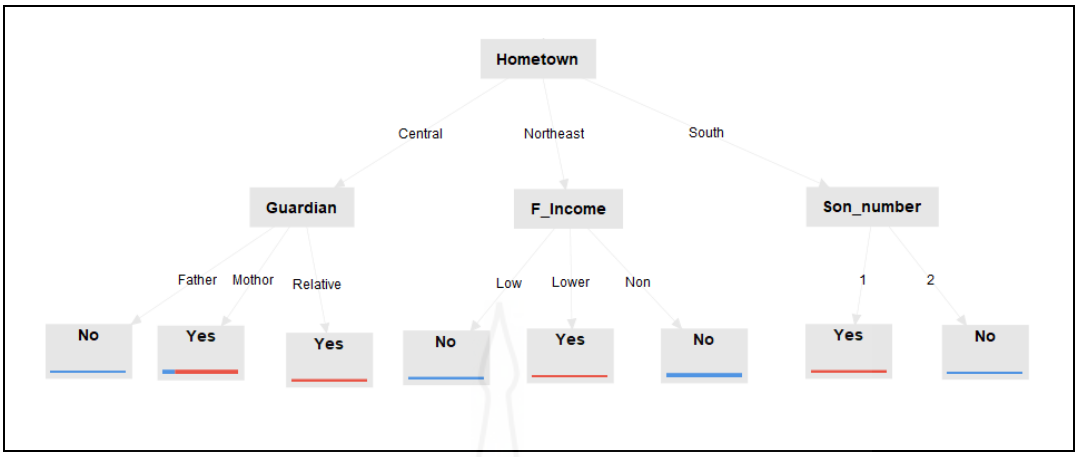


ภาพที่ 4.33 แสดงผลการเปรียบเทียบค่าความถูกต้องโดยใช้ชุดข้อมูล 100, 300 และ 500

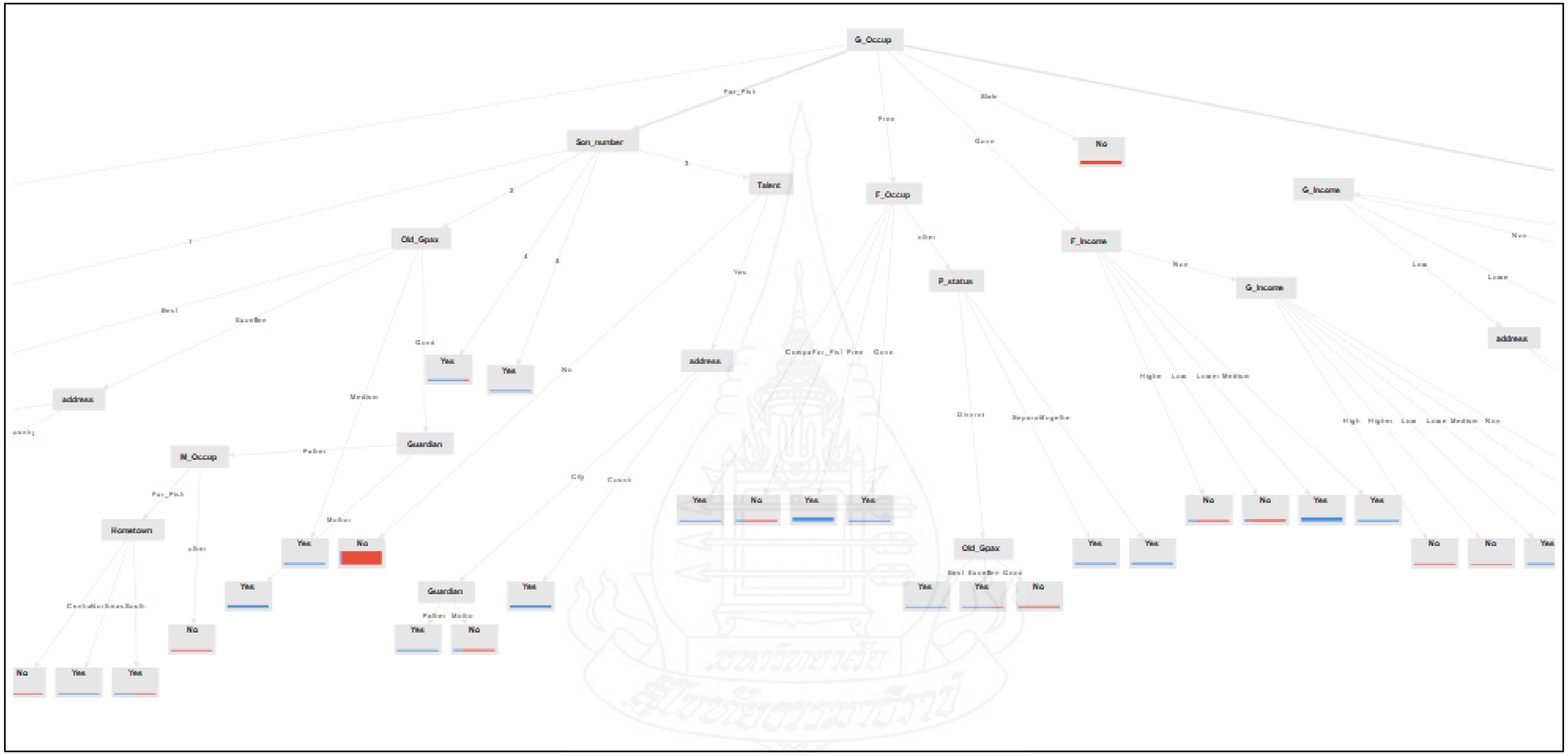
ผู้วิจัยจึงเลือกใช้ชุดข้อมูลทั้งหมดในการสร้างแบบจำลองต้นไม้ต้นสินใจบนพื้นฐานเงื่อนไขหลายเป้าหมาย โดยใช้ชุดข้อมูลทั้งหมด 500 แถว มาดำเนินการแบ่งข้อมูลออกเป็น 2 ส่วน ได้แก่ ข้อมูลฝึกสอน (training data) และข้อมูลทดสอบ (testing data) ออกเป็น 80:20 คือข้อมูลฝึกสอนคิดเป็นร้อยละ 80 หรือจำนวน 400 แถว และข้อมูลทดสอบคิดเป็นร้อยละ 20 หรือจำนวน 100 แถว จากนั้นนำข้อมูลมาสร้างแบบจำลองการจำแนกต้นไม้ตัดสินใจ และใช้วิธี 10-fold cross validation เพื่อให้ข้อมูลมีการกระจายค่าเท่าๆกัน ได้แบบจำลองต้นไม้ตัดสินใจ แสดงดังภาพที่ 4.30 - 4.32



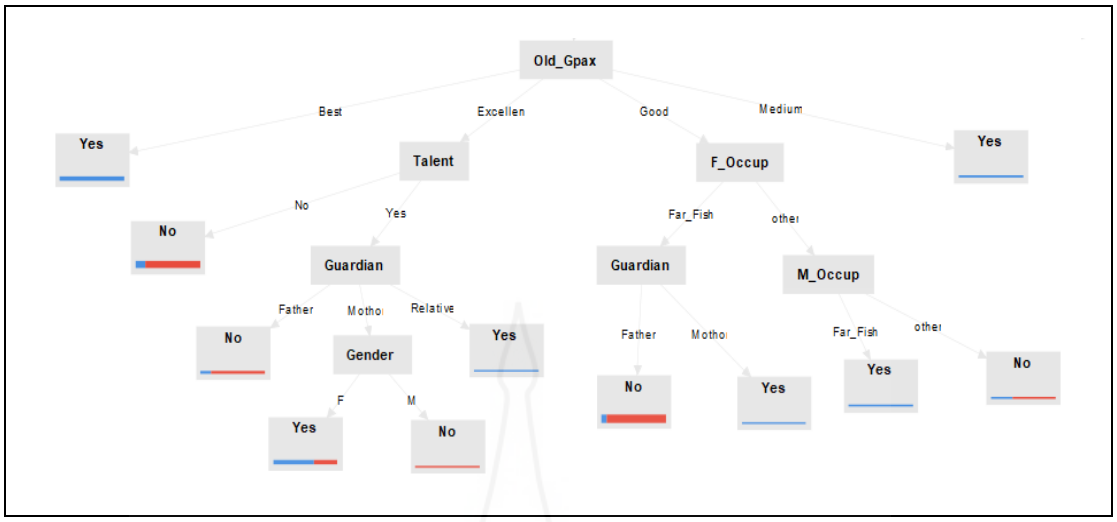
ภาพที่ 4.34 แสดงภาพรวมการสร้างแบบจำลองการจำแนกประเภทต้นไม้ตัดสินใจของเป้าหมาย A



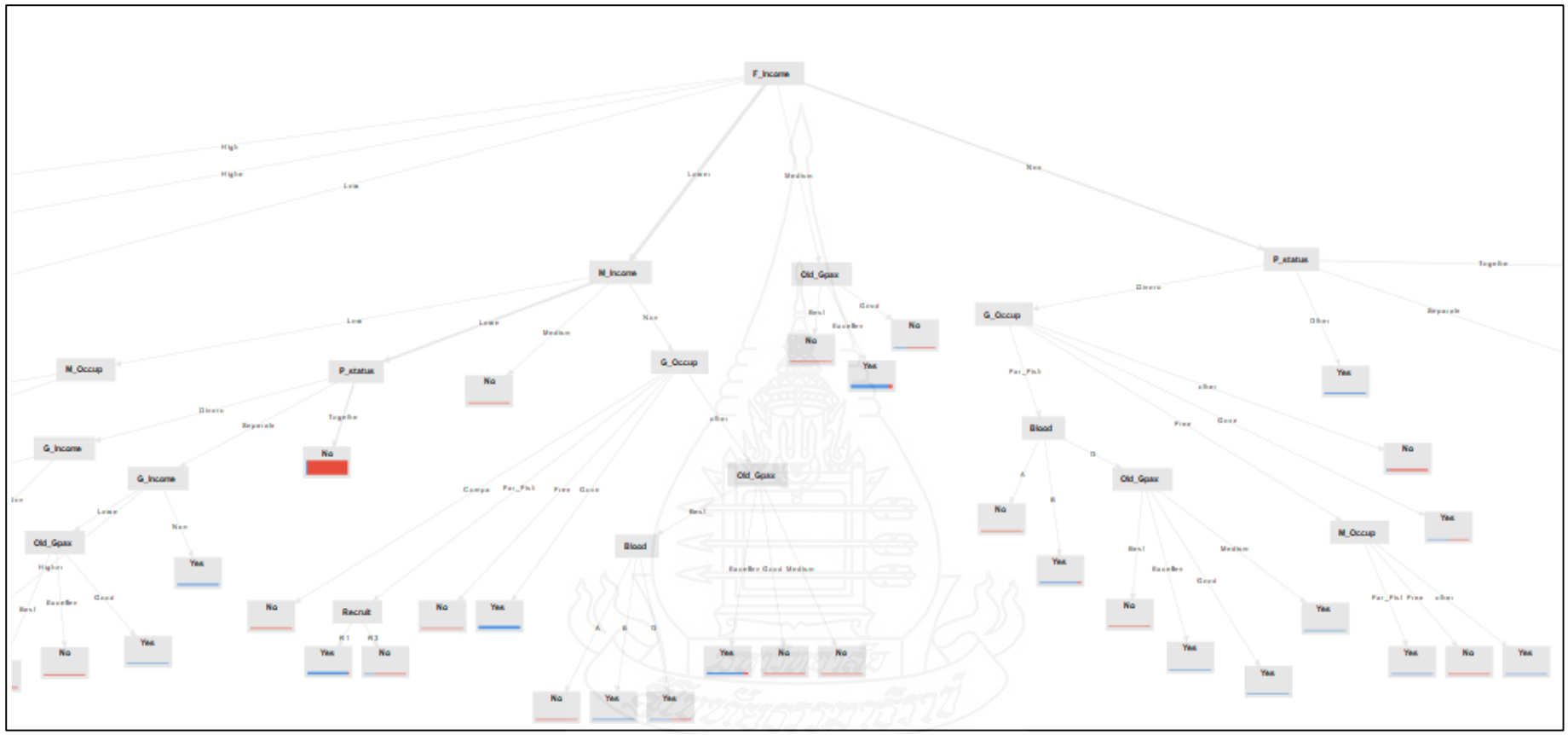
ภาพที่ 4.35 แสดงผลการสร้างแบบจำลองการจำแนกประเภทต้นไม้ตัดสินใจของเป้าหมาย A



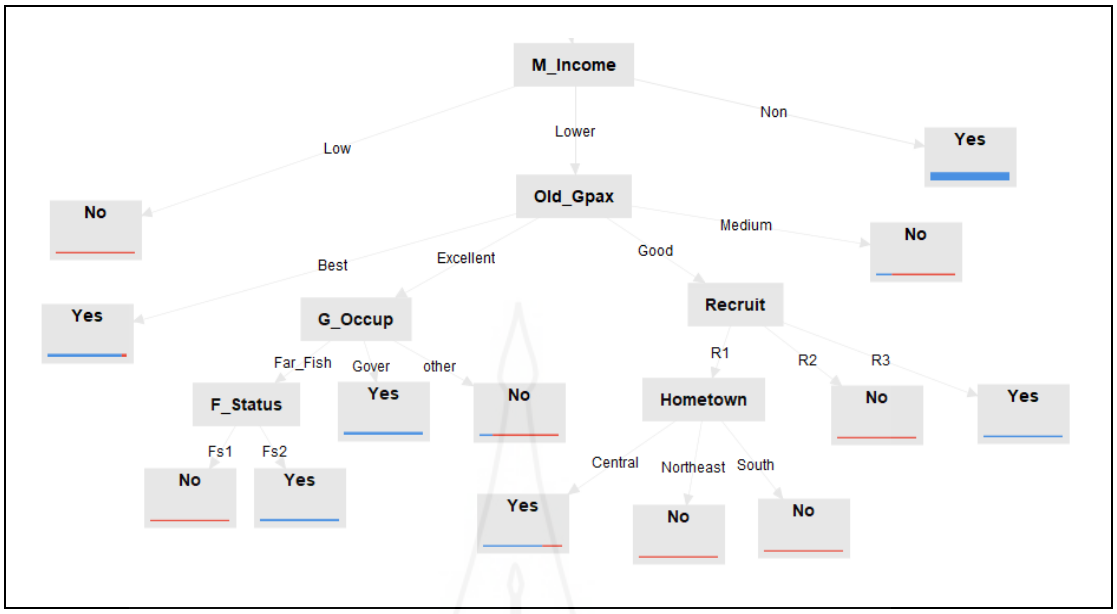
ภาพที่ 4.36 แสดงภาพรวมการสร้างแบบจำลองการจำแนกประเภทต้นไม้ตัดสินใจของเป้าหมาย B



ภาพที่ 4.37 แสดงผลการสร้างแบบจำลองการจำแนกประเภทต้นไม้ตัดสินใจของเป้าหมาย B



ภาพที่ 4.38 แสดงภาพรวมการสร้างแบบจำลองการจำแนกประเภทต้นไม้ตัดสินใจของเป้าหมาย C



ภาพที่ 4.39 แสดงผลการสร้างแบบจำลองการจำแนกประเภทต้นไม้ตัดสินใจของเป้าหมาย C

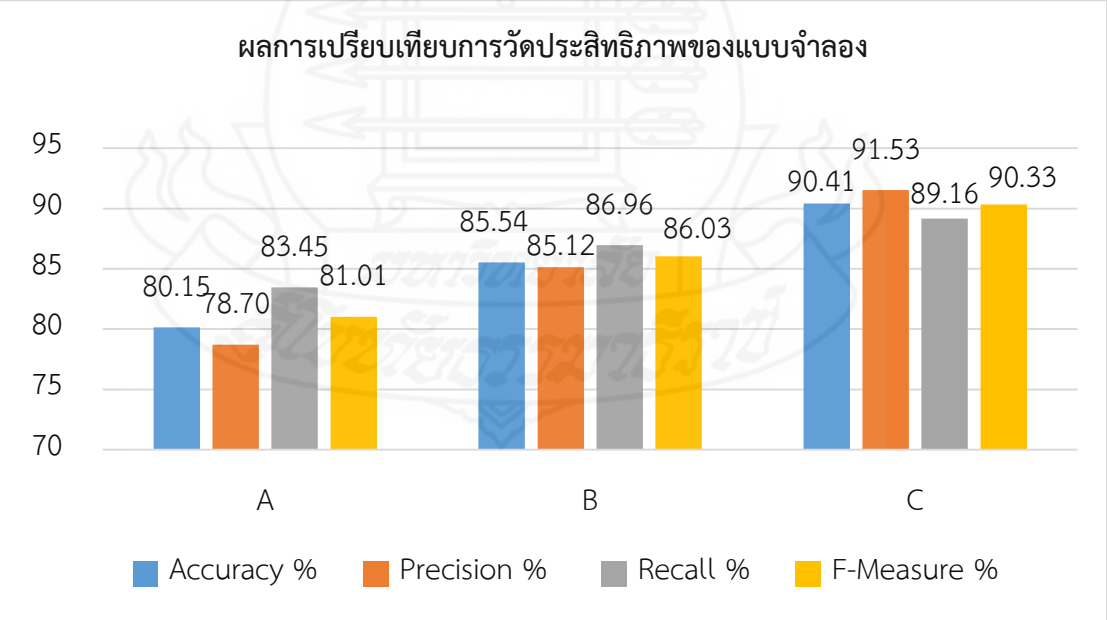
4. การวัดประสิทธิภาพโมเดล (evaluation the model)

งานวิจัยนี้วัดประสิทธิภาพแบบจำลอง โดยพิจารณาจากค่าความถูกต้อง (accuracy) ค่าความแม่นยำ (precision) ค่าเรียกคืน (recall) และค่าประสิทธิภาพโดยรวม (f-measure) ซึ่งเป็นการคำนวณจากตาราง confusion matrix ผลการวัดประสิทธิภาพการเลือกคุณลักษณะสำคัญแบบพลวัตสำหรับความเหมาะสมที่สุดของการจำแนกต้นไม้ตัดสินใจบนพื้นฐานเงื่อนไขหลายเป้าหมายด้วยการเลือกคุณลักษณะที่แปรผันไปตามเป้าหมาย จำนวน 3 เป้าหมาย ได้แก่ 1) เป้าหมาย A 2) เป้าหมาย B และ 3) เป้าหมาย C ได้ผลการทดลองดังตารางที่ 4.22

ตารางที่ 4.22 แสดงผลการวัดประสิทธิภาพของแบบจำลอง

target	accuracy %	precision %	recall %	f-measure %
A	80.15	78.70	83.45	81.01
B	85.54	85.12	86.96	86.03
C	**90.41**	**91.53**	**89.16**	**90.33**
average	**85.37**	**85.12**	**86.52**	**85.79**

จากตารางที่ 4.22 ผลการวัดประสิทธิภาพแบบจำลอง พบว่า แบบจำลองการจำแนกต้นไม้ตัดสินใจบนพื้นฐานเงื่อนไขหลายเป้าหมาย ให้ค่าความถูกต้องร้อยละ 85.37 ค่าความแม่นยำร้อยละ 85.12 ค่าเรียกคืนร้อยละ 86.52 และการวัดประสิทธิภาพโดยรวมร้อยละ 85.79

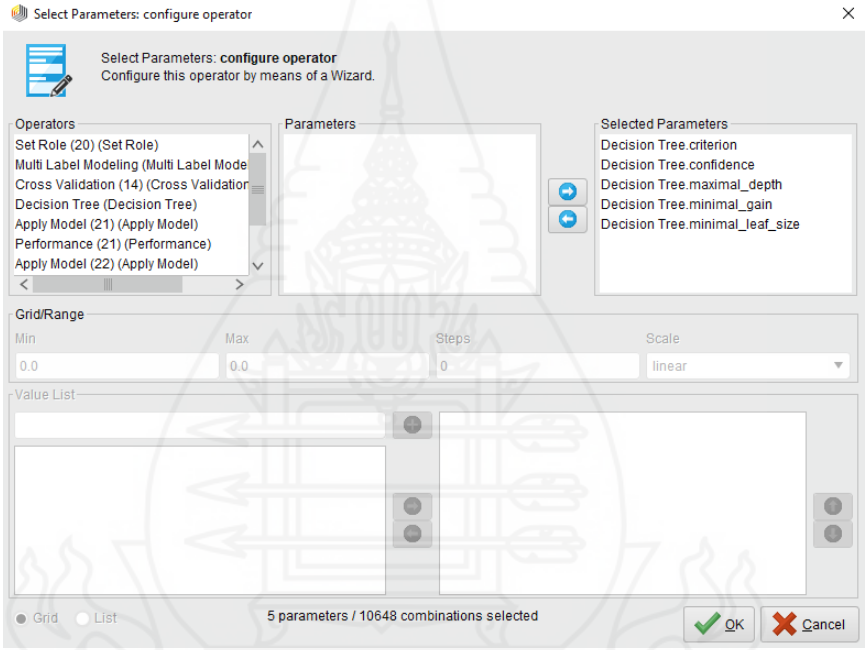


ภาพที่ 4.40 แสดงผลการเปรียบเทียบการวัดประสิทธิภาพของแบบจำลอง

จากภาพที่ 4.40 พบว่า เป้าหมาย C มีค่าความถูกต้องมากที่สุดที่ร้อยละ 90.41 ค่าความแม่นยำร้อยละ 91.53 ค่าเรียกคืนร้อยละ 89.16 และการวัดประสิทธิภาพโดยรวมร้อยละ 90.33

5. การปรับค่าพารามิเตอร์ (parameter tuning)

ทำการปรับค่าพารามิเตอร์ต่าง ๆ โดยใช้ optimize paramiter grids เพื่อให้ได้ค่าที่มีความเหมาะสมกับชุดข้อมูล และแบบจำลองการจำแนกต้นไม้ตัดสินใจบนพื้นฐานเงื่อนไขหลายเป้าหมาย จากเดิมมีการกำหนดค่าพารามิเตอร์ดังนี้ กำหนดค่า criterion เป็นค่า information_gain, ค่า confidence เท่ากับ 0.1, ค่า maximal depth เท่ากับ 10, ค่า minimal gain เท่ากับ 0.01 และกำหนดค่า minimal leaf size เท่ากับ 2 จากนั้นผู้วิจัยได้ทำการปรับแต่งค่าพารามิเตอร์โดยใช้ optimize paramiter grids เพื่อให้ได้ค่าค่าพารามิเตอร์ที่เหมาะสมที่สุดสำหรับชุดข้อมูลและแบบจำลอง โดยทำการกำหนดในส่วนของ selected parameaters ในการกำหนดให้เลือกค่าค่าพารามิเตอร์ต่าง ๆ ที่เหมาะสมที่สุดของแต่ละเป้าหมาย ได้ผลแสดงดังภาพที่ 4.41



ภาพที่ 4.41 แสดงการปรับค่าพารามิเตอร์ของแบบจำลอง

6. การใช้งานการทำนายแบบจำลอง (model prediction)

การใช้งานการทำนายแบบจำลอง ถูกนำไปใช้งานจริงร่วมกับการพิจารณาของคณะกรรมการในการตัดสินใจให้ทุนแก่นักศึกษาของวิทยาลัยฯ อย่างยุติธรรม และโปร่งใส โดยการนำไปใช้งานจริงต้องมีการออกแบบและพัฒนาระบบสารสนเทศ ที่สามารถใช้งานได้สะดวก ประมวลผลได้อย่างรวดเร็ว เพื่อให้สามารถแสดงผลลัพธ์ที่ตอบโจทย์การนำไปใช้ประกอบการพิจารณา ซึ่งทำการแบ่งสัดส่วนน้ำหนักของคะแนนการประเมิน 100 คะแนน แบ่งเป็น 2 ส่วน ได้แก่ คณะกรรมการให้น้ำหนักอยู่ที่ร้อยละ 70 และแบบจำลองให้น้ำหนักอยู่ที่ร้อยละ 30 ทั้งนี้เพื่อลดการเอนเอียง (bias) ในการพิจารณา

บทที่ 5
สรุปผลการวิจัย อภิปรายผล และข้อเสนอแนะ

งานวิจัยนี้เป็นการนำเสนอการเลือกคุณลักษณะสำคัญแบบพลวัตสำหรับความเหมาะสมที่สุดของการจำแนกต้นไม้ตัดสินใจบนพื้นฐานเงื่อนไขหลายเป้าหมาย มีวัตถุประสงค์เพื่อบูรณาการอัลกอริทึมการเลือกคุณลักษณะสำคัญแบบพลวัตกับการจำแนกต้นไม้ตัดสินใจบนพื้นฐานเงื่อนไขหลายเป้าหมาย และเพื่อประเมินประสิทธิภาพการเลือกคุณลักษณะสำคัญแบบพลวัตสำหรับความเหมาะสมที่สุดของการจำแนกต้นไม้ตัดสินใจบนพื้นฐานเงื่อนไขหลายเป้าหมาย มีการดำเนินการประกอบด้วย 6 ขั้นตอนหลัก ได้แก่ 1) การเก็บรวบรวมข้อมูลของนักศึกษาจากงานทุนการศึกษาของวิทยาลัยเทคโนโลยีทางการแพทย์และสาธารณสุข กาญจนาภิเษก ข้อมูลจำนวน 500 คน และคุณลักษณะสำคัญจำนวน 29 คุณลักษณะ 2) การเตรียมข้อมูลโดยกำหนดเงื่อนไขหลายเป้าหมายจำนวน 3 แบบ สำหรับทุนที่มี 3 ประเภท โดยประยุกต์วิธีสังเคราะห์ข้อมูลเพิ่มสำหรับแก้ไขปัญหาชุดข้อมูลของนักศึกษาที่มีความไม่สมดุล และพัฒนาวิธีการเลือกคุณลักษณะสำคัญแบบพลวัตบนพื้นฐานเงื่อนไขหลายเป้าหมาย 3) การสร้างแบบจำลองการจำแนกด้วยอัลกอริทึมต้นไม้ตัดสินใจ สำหรับสอนและทดสอบแบบจำลอง 4) การประเมินประสิทธิภาพแบบจำลอง 5) การปรับค่าพารามิเตอร์เพื่อหาค่าความเหมาะสมที่สุด และ 6) การใช้งานการทำนายแบบจำลอง ได้ผลการทดลอง ดังนี้

1. สรุปการวิจัย

งานวิจัยนี้มีวัตถุประสงค์เพื่อบูรณาการการเลือกคุณลักษณะสำคัญแบบพลวัตกับการจำแนกต้นไม้ตัดสินใจบนพื้นฐานเงื่อนไขหลายเป้าหมาย และเพื่อประเมินประสิทธิภาพการเลือกคุณลักษณะสำคัญแบบพลวัตสำหรับความเหมาะสมที่สุดของการจำแนกต้นไม้ตัดสินใจบนพื้นฐานเงื่อนไขหลายเป้าหมาย จากการวิจัยได้ผลการทดลอง ดังนี้

1. ผลการบูรณาการการเลือกคุณลักษณะสำคัญแบบพลวัตกับการจำแนกต้นไม้ตัดสินใจบนพื้นฐานเงื่อนไขหลายป้าหมาย พบว่า เทคนิค Evolutionary Selection ให้ค่าความถูกต้องดีที่สุด เมื่อนำเทคนิค Evolutionary Selection ไปใช้ในการเลือกคุณลักษณะสำคัญแบบพลวัตบนพื้นฐานเงื่อนไขหลายเป้าหมาย พบว่า สามารถคัดเลือกคุณลักษณะสำคัญของแต่ละเป้าหมายได้แตกต่างกัน โดยสามารถลดจำนวนคุณลักษณะสำคัญจาก 29 คุณลักษณะ ของเป้าหมาย A เหลือจำนวน 27 คุณลักษณะ ส่วนเป้าหมาย B เหลือจำนวน 20 คุณลักษณะ และเป้าหมาย C เหลือจำนวน 20 คุณลักษณะ

2. ผลการประเมินประสิทธิภาพการเลือกคุณลักษณะสำคัญแบบพลวัตสำหรับความเหมาะสมที่สุดของการจำแนกต้นไม้ตัดสินใจบนพื้นฐานเงื่อนไขหลายเป้าหมาย ประสิทธิภาพการพยากรณ์ของแบบจำลองการจำแนกต้นไม้ตัดสินใจบนพื้นฐานเงื่อนไขหลายเป้าหมาย ให้ค่าความถูกต้องร้อยละ

85.37 ค่าความแม่นยำร้อยละ 85.12 ค่าเรียกคืนร้อยละ 86.52 และการวัดประสิทธิภาพโดยรวมร้อยละ 85.79 โดยเป้าหมาย C ได้ค่าความถูกต้องมากที่สุด ซึ่งให้ค่าความถูกต้องร้อยละ 90.41 ค่าความแม่นยำร้อยละ 91.53 ค่าเรียกคืนร้อยละ 89.16 และการวัดประสิทธิภาพโดยรวมร้อยละ 90.33

2. อภิปรายผล

การเลือกคุณลักษณะสำคัญแบบพลวัต เป็นการคัดเลือกคุณลักษณะที่แปรผันไปตามเป้าหมาย เพื่อให้ได้คุณลักษณะสำคัญที่มีผลต่อเป้าหมายการได้รับทุนการศึกษาแต่ละประเภทและเหมาะสมกับแบบจำลองการจำแนกต้นไม้ตัดสินใจ งานวิจัยนี้มีวัตถุประสงค์เพื่อบูรณาการการเลือกคุณลักษณะสำคัญแบบพลวัตกับการจำแนกต้นไม้ตัดสินใจบนพื้นฐานเงื่อนไขหลายเป้าหมาย โดยทำการศึกษาเทคนิค วิธีการแก้ไขปัญหาที่มีผลต่อประสิทธิภาพการพยากรณ์ แต่เนื่องจากชุดข้อมูลที่เก็บรวบรวมมามีความไม่สมดุลของข้อมูล (imbalance data) ด้วยจำนวนการไม่ได้รับทุนมากกว่าการได้รับทุน จึงใช้วิธีการสุ่มตัวอย่างกลุ่มน้อยสังเคราะห์ (SMOTE) สำหรับเพิ่มข้อมูลให้มีจำนวนใกล้เคียงกัน ก่อนนำชุดข้อมูลไปทำการคัดเลือกคุณลักษณะสำคัญ สอดคล้องกับงานวิจัยของ (ภรันยา ปาลวิสุทธิ์, 2559) ที่ใช้เทคนิคการสุ่มตัวอย่างกลุ่มน้อยสังเคราะห์ (SMOTE) มาแก้ไขปัญหาความไม่สมดุลของข้อมูลซึ่งสามารถเพิ่มประสิทธิภาพของแบบจำลองได้ดีขึ้นในทุกเทคนิคที่ได้นำมาเปรียบเทียบ ซึ่งพิจารณาจากค่าความแม่นยำเพิ่มขึ้นเฉลี่ยร้อยละ 5.24 ค่าความไวเพิ่มขึ้นเฉลี่ยร้อยละ 13.82 และค่าความจำเพาะเพิ่มขึ้นเฉลี่ยร้อยละ 8.47 จากนั้นผู้วิจัยนำชุดข้อมูลไปคัดเลือกด้วยเทคนิคการเลือกคุณลักษณะสำคัญ จำนวน 7 เทคนิค เพื่อให้ได้เทคนิคเหมาะสมที่สุดกับแบบจำลองการจำแนกต้นไม้ตัดสินใจบนพื้นฐานเงื่อนไขหลายเป้าหมาย ซึ่งพบว่า เทคนิค Evolutionary Selection ให้ค่าความถูกต้องดีที่สุด สามารถลดจำนวนคุณลักษณะลง เหลือเพียงคุณลักษณะสำคัญที่สอดคล้องกับเป้าหมาย ซึ่งเป็นการสุ่มเลือกคุณลักษณะเข้ามาทีละตัวเพื่อหาค่าประสิทธิภาพในการพยากรณ์คำตอบ หากค่าประสิทธิภาพสูงขึ้นจะเก็บคุณลักษณะนั้นไว้และหากประสิทธิภาพต่ำลงจะถอดคุณลักษณะนั้นออก สอดคล้องกับงานวิจัยของ (อัจจิมา มณฑาพันธุ์, 2562) ทำการวิจัยเรื่อง “การเปรียบเทียบวิธีการคัดเลือกคุณลักษณะที่สำคัญในการปรับปรุงการพยากรณ์มะเร็งเต้านม” พบว่า เทคนิค Evolutionary Selection สามารถลดจำนวนคุณลักษณะสำคัญลงและเพิ่มประสิทธิภาพความแม่นยำในการพยากรณ์ การสร้างแบบจำลองการจำแนกด้วยเทคนิคต้นไม้ตัดสินใจ (decision tree classification) สำหรับฝึกสอนและทดสอบแบบจำลอง และปรับแต่งค่าพารามิเตอร์ให้มีค่าความเหมาะสมที่สุด เพื่อเพิ่มประสิทธิภาพให้กับแบบจำลอง และการประเมินประสิทธิภาพแบบจำลองการจำแนกต้นไม้ตัดสินใจบนพื้นฐานเงื่อนไขหลายเป้าหมาย ผลการวิจัยนี้ได้ผลลัพธ์ค่าความถูกต้องร้อยละ 85.37 ค่าความแม่นยำร้อยละ 85.12 ค่าเรียกคืนร้อยละ 86.52 และการวัดประสิทธิภาพโดยรวมร้อยละ 85.79

3. ข้อเสนอแนะ

3.1 ข้อเสนอแนะทั่วไป

1. สำหรับการคัดเลือกคุณลักษณะแบบพลวัตบนพื้นฐานเงื่อนไขหลายเป้าหมาย แสดงให้เห็นว่าการเลือกใช้เทคนิคการคัดเลือกคุณลักษณะที่เหมาะสมขึ้นอยู่กับเป้าหมายหรือผลลัพธ์ที่ต้องการ ดังนั้นจึงจำเป็นต้องทำการเปรียบเทียบเทคนิคการคัดเลือกคุณลักษณะสำคัญที่หลากหลายเพื่อให้ได้เทคนิคที่เหมาะสมที่สุดสำหรับนำไปใช้ในการคัดเลือกคุณลักษณะ

2. การเลือกใช้อัลกอริทึมจำแนกประเภทข้อมูลจะต้องมีความเหมาะสมกับชุดข้อมูล ซึ่งสามารถทดลองปรับเปลี่ยนอัลกอริทึมเพื่อเลือกใช้อัลกอริทึมที่เหมาะสมที่สุดกับชุดข้อมูลการจำแนกประเภท

3.2 ข้อเสนอแนะในการวิจัยครั้งต่อไป

1. การประยุกต์ใช้การเรียนรู้ของเครื่องมีหลายปัจจัยที่มีผลต่อความถูกต้องและแม่นยำ เช่น จำนวนคุณลักษณะพิเศษของข้อมูล รูปแบบของข้อมูล ปริมาณข้อมูล การเลือกแบบจำลอง การกำหนดวิธีการปรับแต่งแบบจำลอง และการประเมินประสิทธิภาพ

2. งานวิจัยนี้สามารถนำไปพัฒนาเพิ่มเติมหรือต่อยอดในกระบวนการปรับแต่งค่าแบบจำลองของเรียนรู้ของเครื่องเพื่อให้ได้ค่าที่เหมาะสมที่สุด รวมถึงการเพิ่มคุณลักษณะของข้อมูลที่มากขึ้น และการปรับเปลี่ยนเป้าหมายหรือเพิ่มเติมทุนการศึกษาอื่นที่ต้องการพยากรณ์ เพื่อให้ได้แบบจำลองการพยากรณ์ที่ได้มีความน่าเชื่อถือมากยิ่งขึ้น

3. สามารถพัฒนาเพิ่มเติมโดยการออกแบบพัฒนาระบบหรือเว็บแอปพลิเคชันสำหรับนำเข้าข้อมูลใหม่ ๆ เพื่อให้สามารถนำไปประยุกต์ใช้งาน โดยทำการทดสอบการทำนายเพื่อนำข้อมูลผลลัพธ์ที่ได้ไปประกอบการพิจารณาตัดสินใจ

บรรณานุกรม

กาญจน์ ณ ศรีธะ กิตติศักดิ์ เกิดประสพ และนิตยา เกิดประสพ. (2561). การเปรียบเทียบเทคนิคการสุ่มตัวอย่างเพื่อจำแนกข้อมูลที่ไม่สมดุล. *วารสารวิทยาการสารสนเทศและเทคโนโลยีประยุกต์*, 1(1), 20-37.

ทรงศักดิ์ ภูสีอ่อน. (2554). *การประยุกต์ใช้ SPSS วิเคราะห์ข้อมูลงานวิจัย*. มหาวิทยาลัยมหาสารคาม, มหาสารคาม.

นิภาพร ชนะมาร และพรรณี สิทธิเดช. (2557, กรกฎาคม-ธันวาคม). การวิเคราะห์ปัจจัยการเรียนรู้ด้วยการคัดเลือกคุณสมบัติและการพยากรณ์. *วารสารมหาวิทยาลัยราชภัฏสกลนคร*, 6(12), 31-46.

พุทธิพร ธนธรรมเมธี และเยาวเรศ ศิริสถิตย์กุล. (2562, พฤศจิกายน-ธันวาคม). เทคนิคการจำแนกข้อมูลที่พัฒนาสำหรับชุดข้อมูลที่ไม่สมดุลของภาวะข้อเข่าเสื่อมในผู้สูงอายุ. *วารสารวิทยาศาสตร์และเทคโนโลยี,* 27(6), 1164-1178.

ภรันยา ปาลวิสุทธิ์. (2559, มกราคม-มิถุนายน). การเพิ่มประสิทธิภาพเทคนิคต้นไม้ตัดสินใจบนชุดข้อมูลที่ไม่สมดุล โดยวิธีการสุ่มเพิ่มตัวอย่างกลุ่มน้อยสำหรับข้อมูลการเป็นโรคติดอินเทอร์เน็ต. *วารสารเทคโนโลยีสารสนเทศ*, 12(1), 54-63.

ภัทราวุฒิ แสงศิริ. (2553, เมษายน-มิถุนายน). การคัดแยกประเภทของมะเร็งเม็ดเลือดขาวโดยใช้วิธีการจัดอันดับร่วมกับเทคนิคซัพพอร์ตเวกเตอร์แมชชีน. *วารสารวิจัย มข. (บศ.)* 10(2), 10-7.

รัชพล กลัดชื่น และจรัญ แสนราช. (2561). การเปรียบเทียบประสิทธิภาพอัลกอริทึมและการคัดเลือกคุณลักษณะที่เหมาะสมเพื่อการทำนายผลสัมฤทธิ์ทางการเรียนของนักศึกษาระดับอาชีวศึกษา. *วารสารวิจัยมหาวิทยาลัยเทคโนโลยีราชมงคลธัญบุรี*. 17(1). 1-10

วฤษาย์ ร่มสายหยุด. (2564). *การเรียนรู้ของเครื่องสำหรับการวิเคราะห์ข้อมูลเชิงทำนายและการประยุกต์.* มหาวิทยาลัยสุโขทัยธรรมาธิราช, นนทบุรี. (320 หน้า).

วิชญ์วิสิฐ เกษรสิทธิ์ จิราวัลย์ จิตรถเวช และวิชิต หล่อจีระชุณห์กุล. (2563, มกราคม). การลดจำนวนกลุ่มในการจำแนกแบบหลายกลุ่มเป็นสองกลุ่มสำหรับการจำแนกการกลับมารักษาซ้ำในโรงพยาบาลของผู้ป่วยโรคเบาหวาน. *วารสารวิทยาศาสตร์และเทคโนโลยี*. 28(1). 41-51

ปราโมทย์ ลือนาม. (2561). "หลักการและวิธีการวิเคราะห์ข้อมูลขนาดใหญ่" ใน *ประมวลสาระชุดวิชาการวิเคราะห์ข้อมูลขนาดใหญ่สำหรับธุรกิจ หน่วยที่ 8-15.* มหาวิทยาลัยสุโขทัยธรรมาธิราช, นนทบุรี.

อัจจิมา มณฑาพันธุ์. (2562, พฤษภาคม-สิงหาคม). การเปรียบเทียบวิธีการคัดเลือกคุณลักษณะที่สำคัญในการปรับปรุงการพยากรณ์มะเร็งเต้านม. *วารสารแพทยสารทหารอากาศ.* 2(65), 49-56.

เอกสิทธิ์ พัชรวงศ์ศักดา. (2557). *การวิเคราะห์ข้อมูลด้วยเทคนิค ดาต้า ไมน์นิ่ง เบื้องต้น (An Introduction to Data Mining Techniques)*. กรุงเทพฯ, 53-7.

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. J. Artif. Int. Res., 16(1), 321-357.

Ethem, A. (2020). Introduction to Machine Learning. *4th Ed. The MIT Press Cambridge.* Massachusetts London: England.

Everitt, B. S. (2010). *Multivariable Modeling and Multivariate Analysis for The Behavioral Sciences*. Taylor & Francis Group, LLC.

He, H., and Ghodsi, A. (2010). Rare class classification by support vector machine. *In Pattern Recognition (ICPR), 2010 20th International Conference on IEEE*, 548-551.

Kaewchinporn, C. (2010). Data Classification with Decision Tree and Clustering Techniques. *Thesis in Computer Science, King Mongkut's Institute of Technology Ladkrabang, Thailand*.

Koller, D. and Mehran, S. (1996). *Toward Optimize Feature Selection. In, 284-92.* Moregan Kaufmann.

Mohd, F. K., Gaurav, C., & Jaitly, A. K. (2011), "An approach to overcome imbalance datasets of eukaryotic genomes during the analysis by machine learning technique (SVM)", *Indian Journal of Science and Technology*, Vol. 4, No. 5, pp. 520-524.

Prajapati. V. (2013). *Big data Analytics with R and Hadoop.* Packt Pubishing.

Romsaiyud, W. Schnoor, H. and Hasselbring, W. (2019). Improving k-Nearest Neighbor Pattern Recognition Models for Privacy-Preserving Data Analysis,". *2019 IEEE International Conference on Big Data (Big Data)*, Los Angeles, CA, USA, 5804-5813, doi: 10.1109/BigData47090.2019.9006281.

Tanaka, E. A., Nozawa, S. R., Macedo, A. A., & Baranauskas, J. A. (2015). A multi-label approach using binary relevance and decision trees applied to functional genomics. Journal of Biomedical Informatics, 54, 85–95. doi:10.1016/j.jbi.2014.12.011

ภาคผนวก

ภาคผนวก ก

หนังสือขอความอนุเคราะห์ให้นักศึกษาเก็บข้อมูลเพื่อการศึกษาวิทยานิพนธ์

ที่ อว ๐๖๐๒.๒๕/๖๙๗

มหาวิทยาลัยสุโขทัยธรรมาธิราช
ตำบลบางพูด อำเภอปากเกร็ด
จังหวัดนนทบุรี ๑๑๑๒๐

๑๗ มิถุนายน ๒๕๖๓

เรื่อง ขอความอนุเคราะห์ให้นักศึกษาเก็บข้อมูลเพื่อการศึกษาวิทยานิพนธ์

เรียน ผู้อำนวยการวิทยาลัยเทคโนโลยีทางการแพทย์และสาธารณสุข กาญจนาภิเษก

สิ่งที่ส่งมาด้วย โครงร่างวิทยานิพนธ์ และรายละเอียดการเก็บข้อมูล จำนวน ๑ ชุด

ด้วยนางสาวนิชาภา จำปาศรี นักศึกษาหลักสูตรวิทยาศาสตรมหาบัณฑิต แขนงวิชาเทคโนโลยีสารสนเทศและการสื่อสาร สาขาวิชาวิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยสุโขทัยธรรมาธิราช กำลังทำการศึกษาวิทยานิพนธ์ เรื่อง "การเลือกคุณลักษณะสำคัญแบบพลวัตสำหรับความเหมาะสมที่สุดของการจำแนกต้นไม้ตัดสินใจบนพื้นฐานเงื่อนไขหลายเป้าหมาย (Dynamic Feature Selection for Optimization of Decision Tree Classification Based on Multi-Target Conditions)" อยู่ในความดูแลของรองศาสตราจารย์ ดร.วฤษาย์ ร่มสายหยุด อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก และอาจารย์ ดร.เอกสิทธิ์ พัชรวงศ์ศักดา อาจารย์ที่ปรึกษาร่วม

ในการศึกษาวิทยานิพนธ์ครั้งนี้ นักศึกษามีความประสงค์ขอเก็บข้อมูลนักศึกษาวิทยาลัยเทคโนโลยีทางการแพทย์และสาธารณสุข กาญจนาภิเษก ประกอบด้วยข้อมูลคุณลักษณะสำคัญจำนวน ๓๒ คุณลักษณะสำคัญ แถวข้อมูลจำนวน ๕๐๐ แถว เพื่อประกอบการทำการศึกษาวิทยานิพนธ์

จึงเรียนมาเพื่อขอความอนุเคราะห์จากท่าน โปรดพิจารณาอนุญาตให้นักศึกษาดำเนินการเก็บข้อมูลเพื่อประกอบการศึกษาวิทยานิพนธ์ รายละเอียดที่นักศึกษาเสนอมาพร้อมนี้ ตามที่เห็นสมควรด้วยจะขอบคุณยิ่ง

ขอแสดงความนับถือ

(อาจารย์ ดร.สิทธิชัย รัชยศโยธิน)
ประธานกรรมการประจำสาขาวิชาวิทยาศาสตร์และเทคโนโลยี

สาขาวิชาวิทยาศาสตร์และเทคโนโลยี
โทรศัพท์ ๐ ๒๕๐๔ ๘๑๙๓
โทรสาร ๐ ๒๕๐๓ ๔๙๓๒

ภาคผนวก ข

เอกสารรับรองโครงการวิจัย

เอกสารรับรองโครงการวิจัย
โดยคณะกรรมการพิจารณาจริยธรรมการวิจัยในมนุษย์
วิทยาลัยเทคโนโลยีทางการแพทย์และสาธารณสุข กาญจนาภิเษก

เอกสารรับรองเลขที่ : KMPHT- 63010011

ชื่อโครงการ/งานวิจัย : การเลือกคุณลักษณะสำคัญแบบพลวัตสำหรับความเหมาะสมที่สุดของการจำแนกต้นไม้ตัดสินใจบนพื้นฐานเงื่อนไขหลายเป้าหมาย

ชื่อผู้ดำเนินการวิจัย : นางสาวนิชาภา จำปาศรี

เอกสารรับรอง :

1. แบบเสนอโครงการวิจัย
2. เอกสารชี้แจงผู้เข้าร่วมการวิจัย
3. หนังสือยินยอมตนให้ทำการวิจัย
4. แบบเก็บรวบรวมข้อมูล/โปรแกรมหรือกิจกรรม

วันที่รับรอง : 1 กรกฎาคม 2563

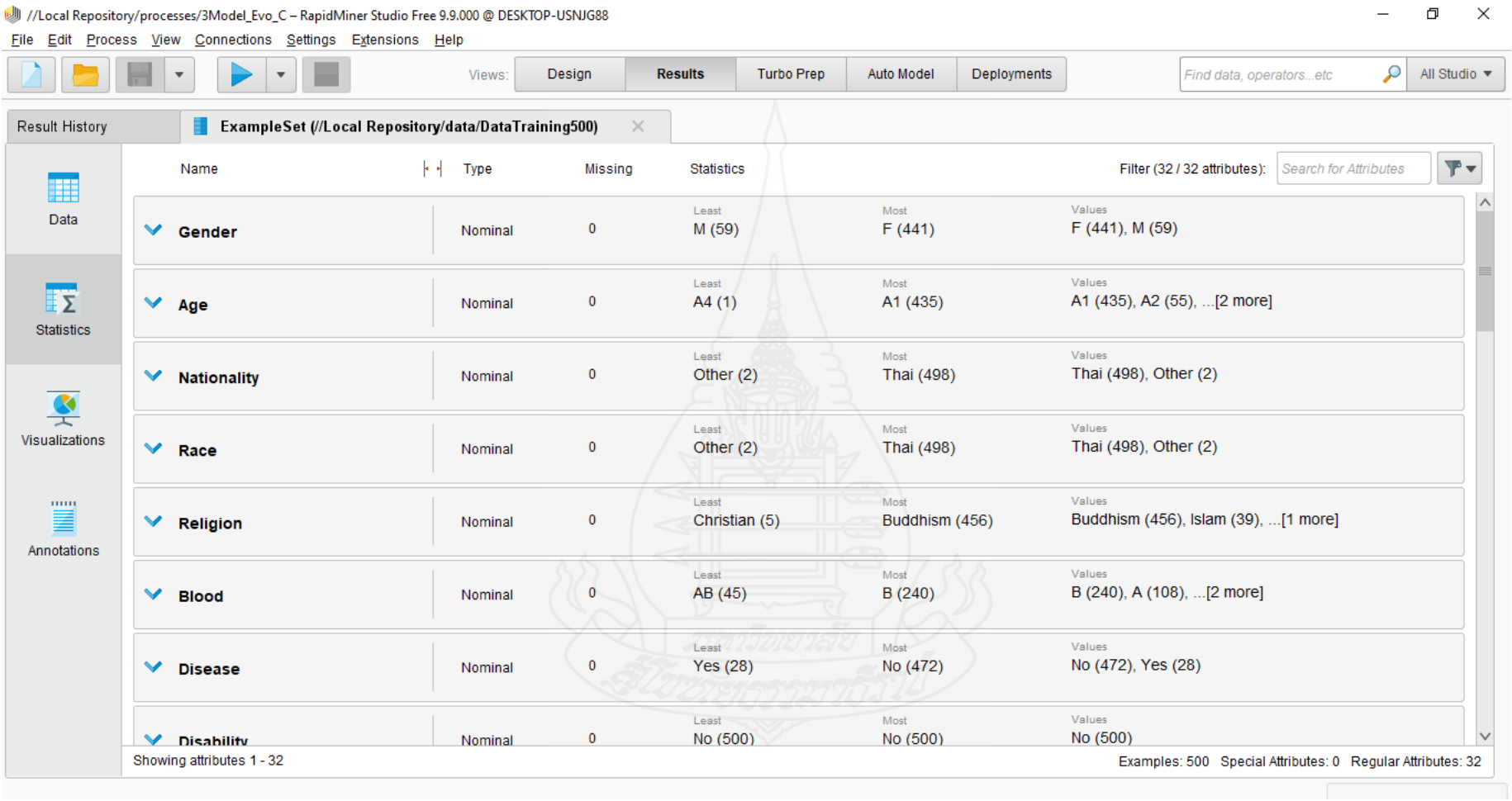
วันหมดอายุ : 30 มิถุนายน 2564

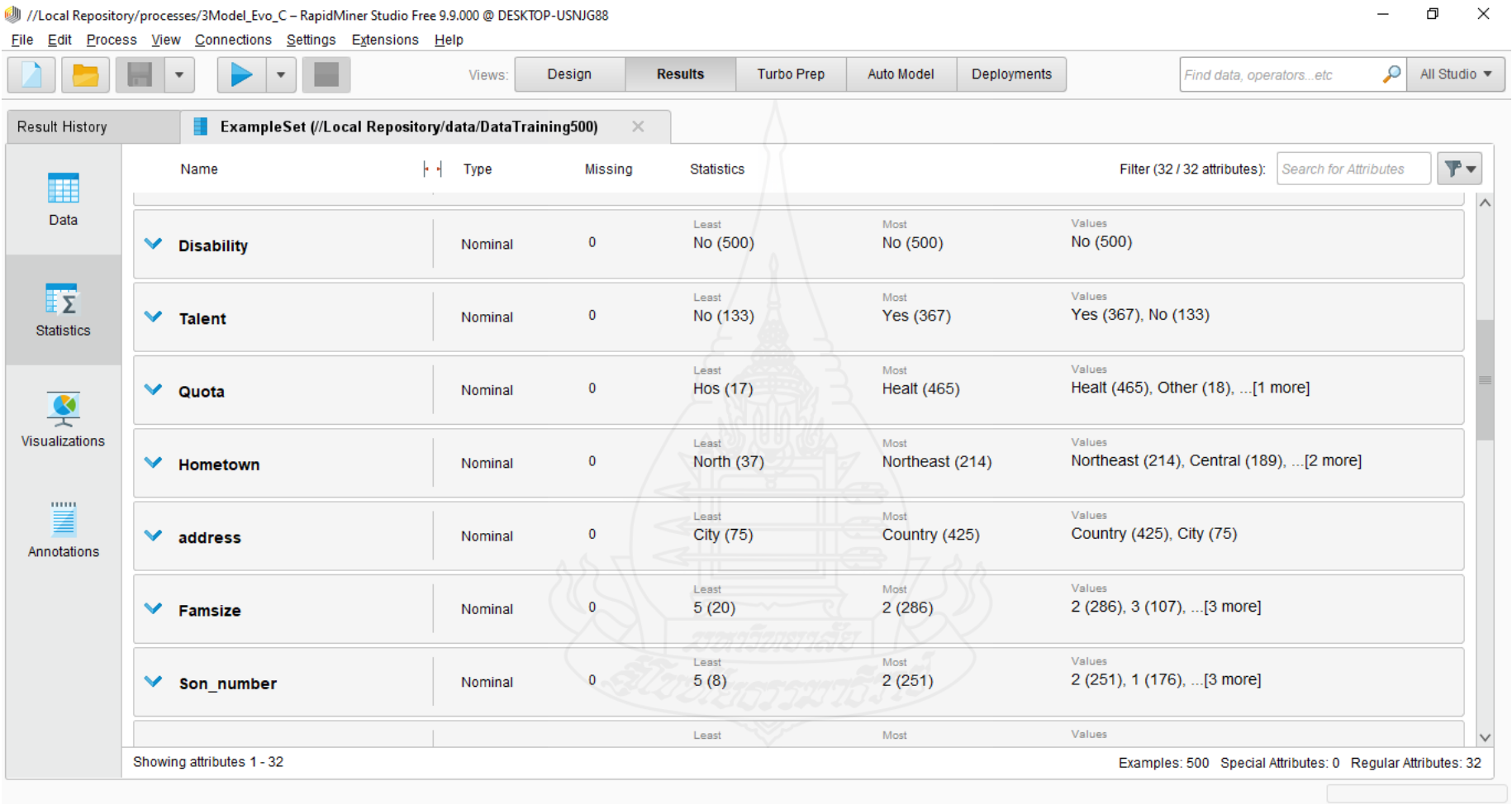
ขอรับรองว่าโครงการวิจัยดังกล่าวข้างต้นได้ผ่านการพิจารณาเห็นชอบให้ดำเนินการ โดยสอดคล้องกับคำประกาศเฮลซิงกิ จากคณะกรรมการพิจารณาจริยธรรมการวิจัยในมนุษย์ของวิทยาลัยเทคโนโลยีทางการแพทย์และสาธารณสุข กาญจนาภิเษก

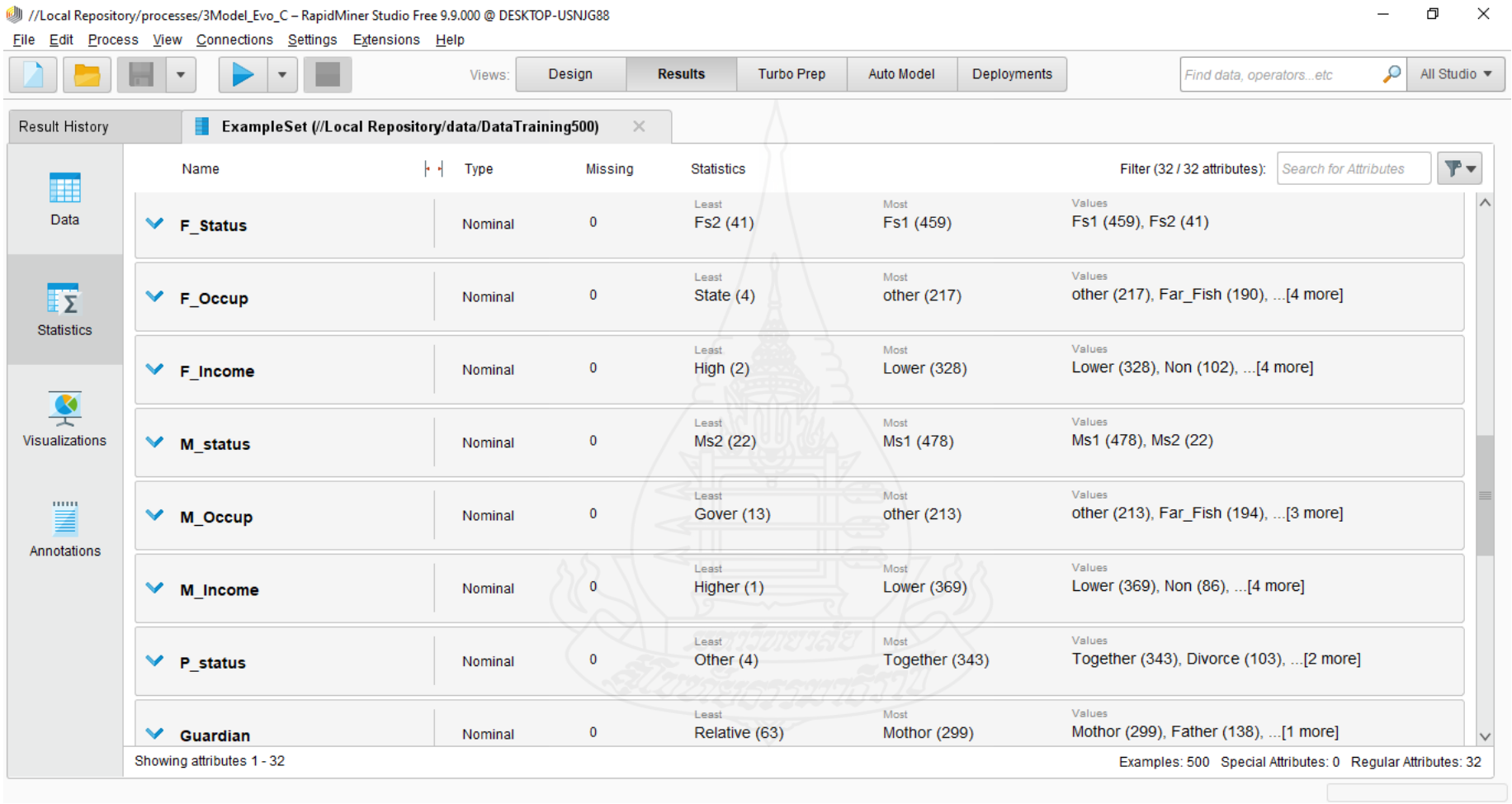
(ดร.นพมาส เครือสุวรรณ)
วิทยาจารย์ชำนาญการพิเศษ
ประธานกรรมการพิจารณาจริยธรรมการวิจัยในมนุษย์
วิทยาลัยเทคโนโลยีทางการแพทย์และสาธารณสุข กาญจนาภิเษก

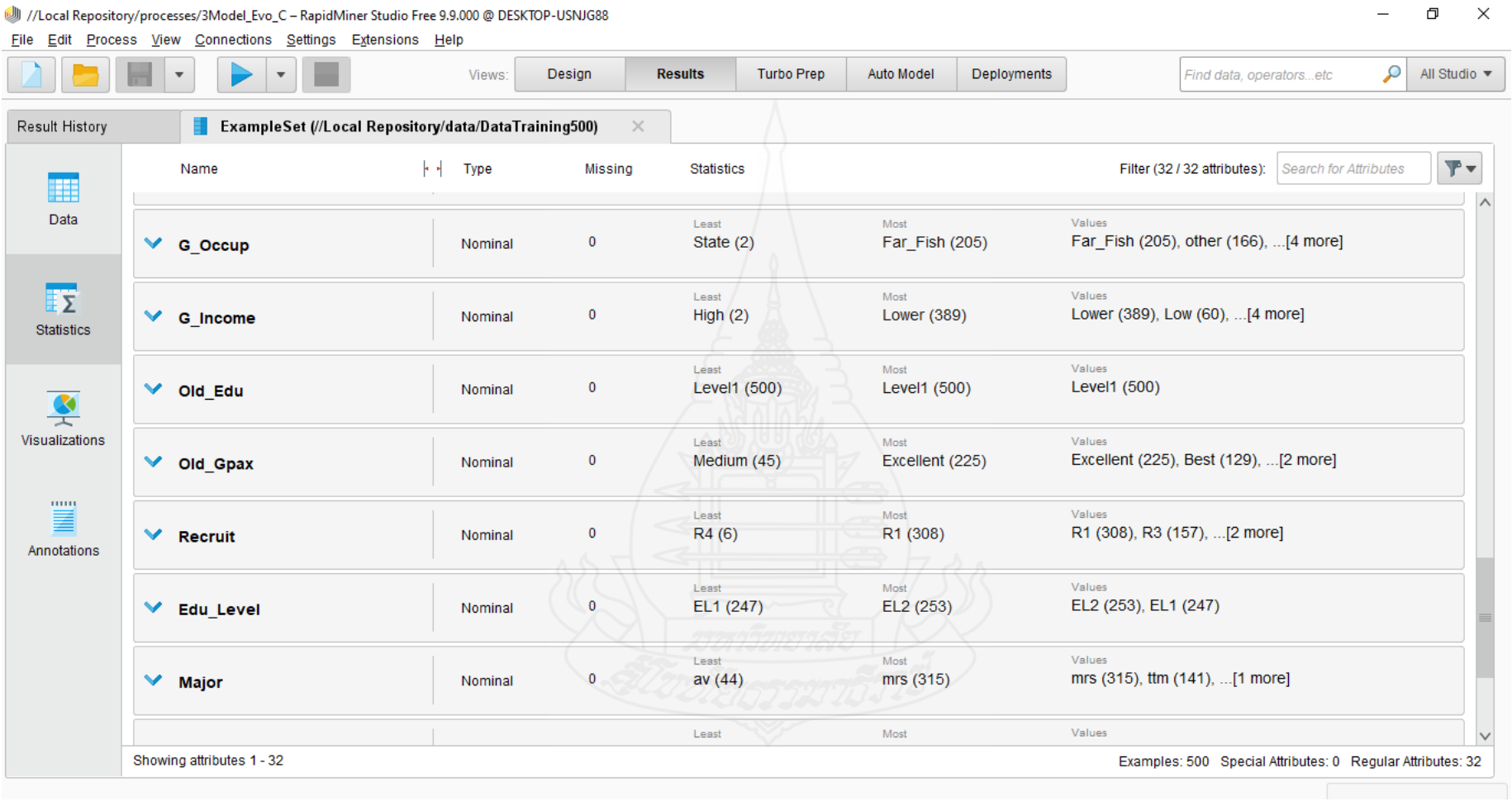
ภาคผนวก ค

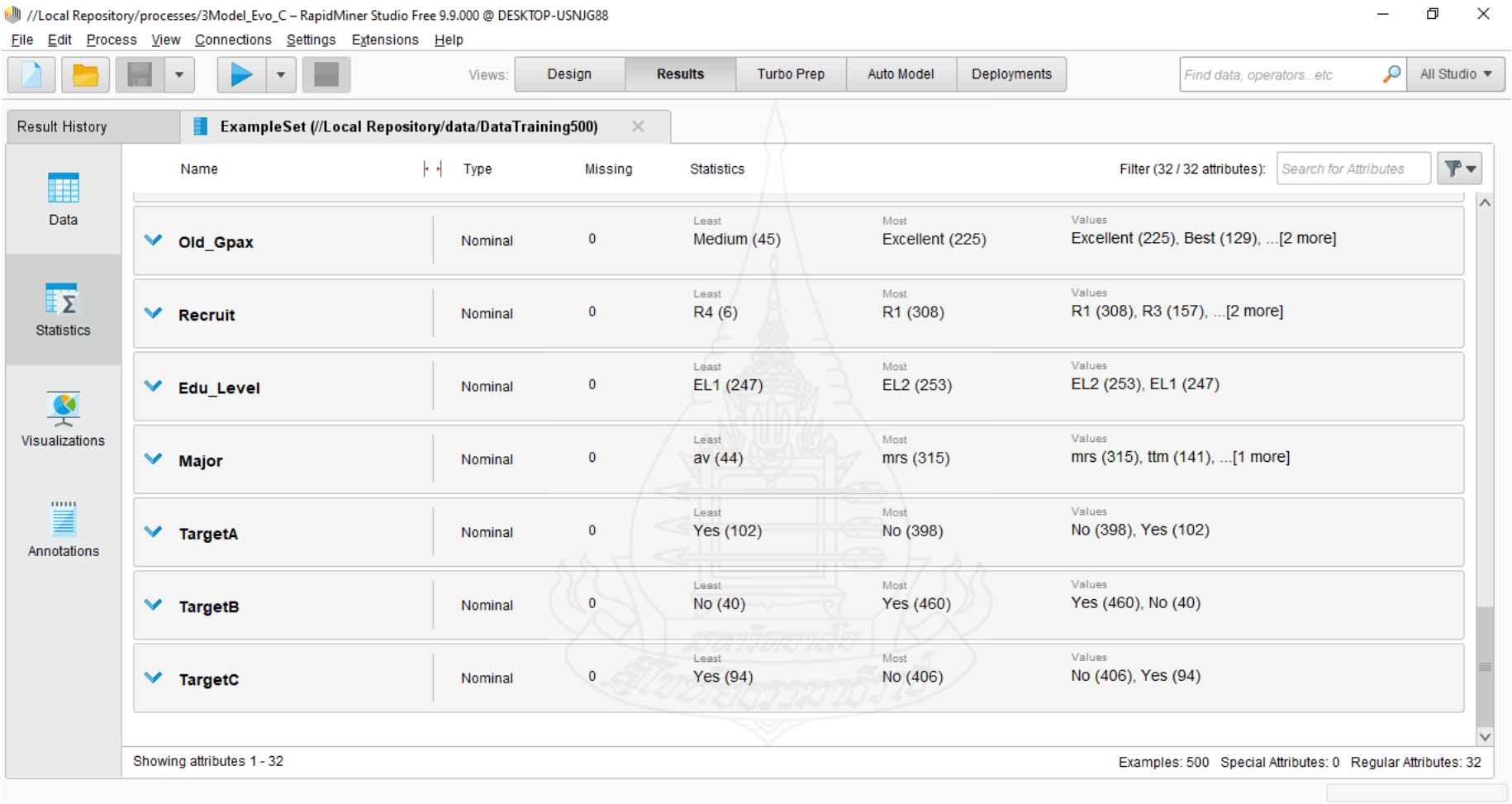
ชุดข้อมูล (dataset)

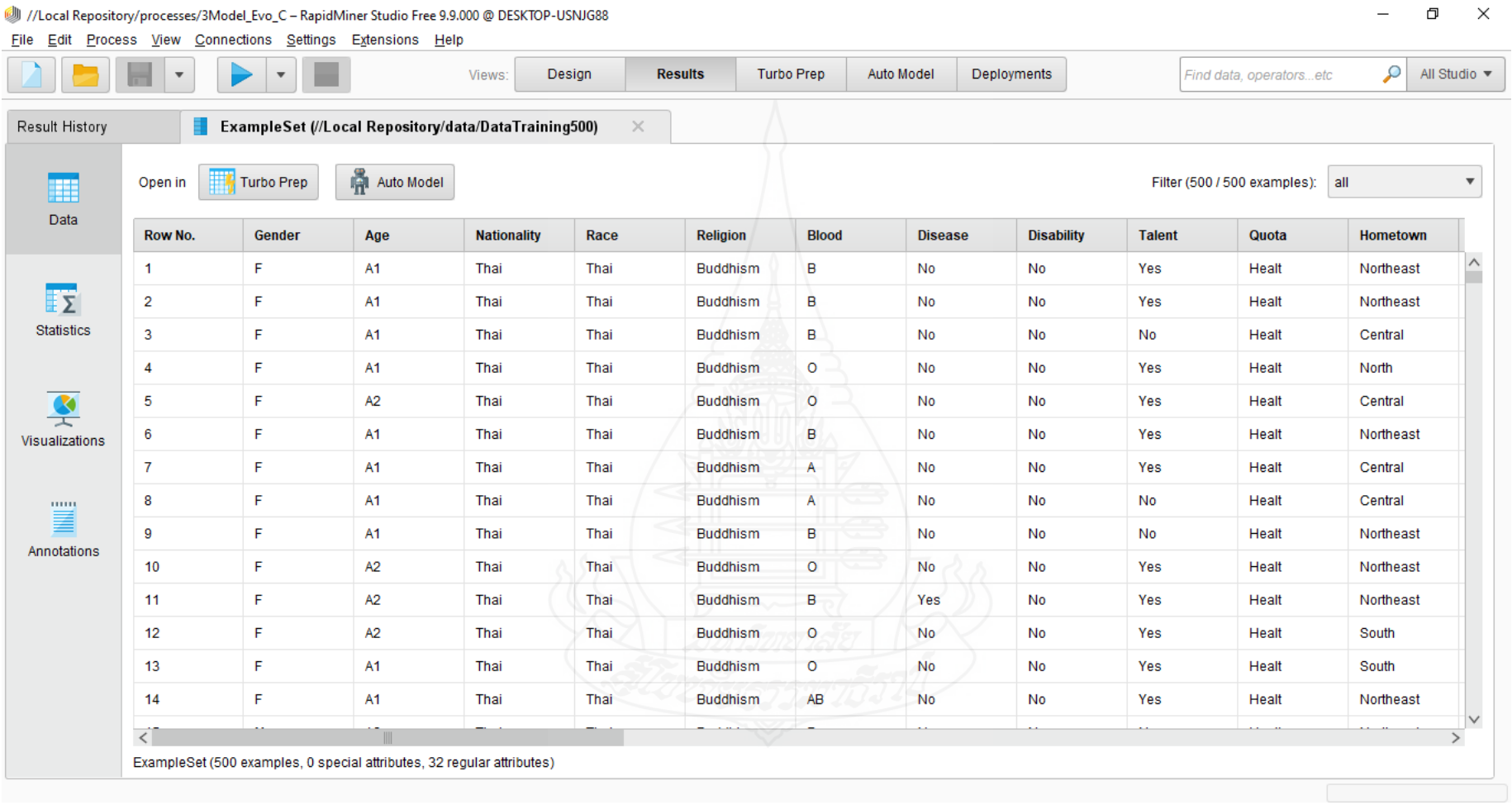

//Local Repository/processes/3Model_Evo_C – RapidMiner Studio Free 9.9.000 @ DESKTOP-USNJG88
File Edit Process View Connections Settings Extensions Help
Views: Design Results Turbo Prep Auto Model Deployments
Find data, operators...etc
All Studio
Result History
ExampleSet (//Local Repository/data/DataTraining500)
Data
Statistics
Visualizations
Annotations
Name Type Missing Statistics
Filter (32 / 32 attributes): Search for Attributes
Gender Nominal 0 Least M (59) Most F (441) Values F (441), M (59)
Age Nominal 0 Least A4 (1) Most A1 (435) Values A1 (435), A2 (55), ...[2 more]
Nationality Nominal 0 Least Other (2) Most Thai (498) Values Thai (498), Other (2)
Race Nominal 0 Least Other (2) Most Thai (498) Values Thai (498), Other (2)
Religion Nominal 0 Least Christian (5) Most Buddhism (456) Values Buddhism (456), Islam (39), ...[1 more]
Blood Nominal 0 Least AB (45) Most B (240) Values B (240), A (108), ...[2 more]
Disease Nominal 0 Least Yes (28) Most No (472) Values No (472), Yes (28)
Disability Nominal 0 Least No (500) Most No (500) Values No (500)
Showing attributes 1 - 32
Examples: 500 Special Attributes: 0 Regular Attributes: 32


//Local Repository/processes/3Model_Evo_C – RapidMiner Studio Free 9.9.000 @ DESKTOP-USNJG88
File Edit Process View Connections Settings Extensions Help
Views: Design Results Turbo Prep Auto Model Deployments
Find data, operators...etc
All Studio
Result History
ExampleSet (//Local Repository/data/DataTraining500)
Data
Statistics
Visualizations
Annotations
Name Type Missing Statistics
Filter (32 / 32 attributes):
Search for Attributes
Disability Nominal 0 Least No (500) Most No (500) Values No (500)
Talent Nominal 0 Least No (133) Most Yes (367) Values Yes (367), No (133)
Quota Nominal 0 Least Hos (17) Most Healt (465) Values Healt (465), Other (18), ...[1 more]
Hometown Nominal 0 Least North (37) Most Northeast (214) Values Northeast (214), Central (189), ...[2 more]
address Nominal 0 Least City (75) Most Country (425) Values Country (425), City (75)
Famsize Nominal 0 Least 5 (20) Most 2 (286) Values 2 (286), 3 (107), ...[3 more]
Son_number Nominal 0 Least 5 (8) Most 2 (251) Values 2 (251), 1 (176), ...[3 more]
Least Most Values
Showing attributes 1 - 32
Examples: 500 Special Attributes: 0 Regular Attributes: 32

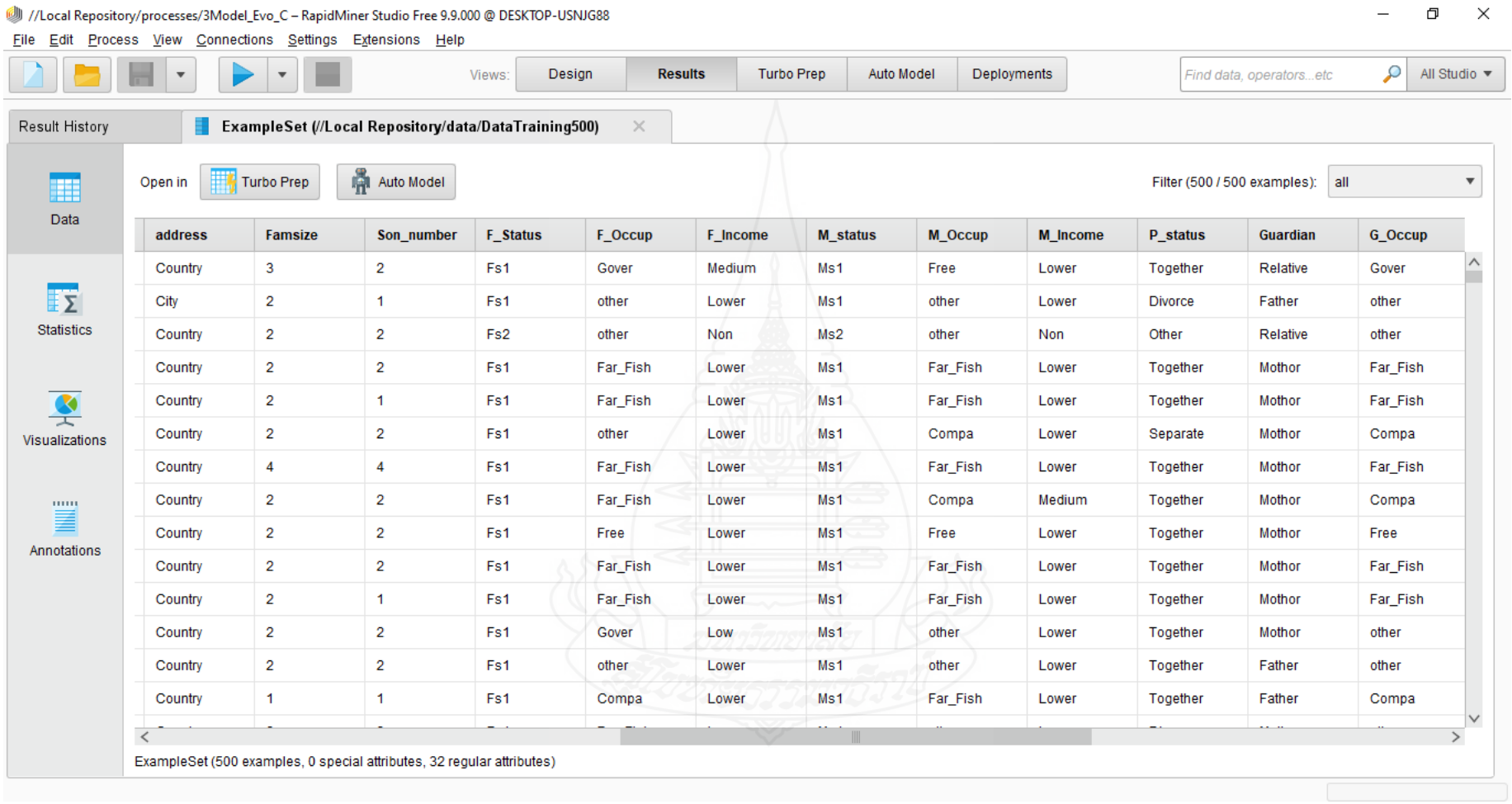

//Local Repository/processes/3Model_Evo_C – RapidMiner Studio Free 9.9.000 @ DESKTOP-USNJG88
File Edit Process View Connections Settings Extensions Help
Views: Design Results Turbo Prep Auto Model Deployments
Find data, operators...etc
All Studio
Result History
ExampleSet (//Local Repository/data/DataTraining500)
Data
Statistics
Visualizations
Annotations
Name Type Missing Statistics
Filter (32 / 32 attributes): Search for Attributes
F_Status Nominal 0 Least Fs2 (41) Most Fs1 (459) Values Fs1 (459), Fs2 (41)
F_Occup Nominal 0 Least State (4) Most other (217) Values other (217), Far_Fish (190), ...[4 more]
F_Income Nominal 0 Least High (2) Most Lower (328) Values Lower (328), Non (102), ...[4 more]
M_status Nominal 0 Least Ms2 (22) Most Ms1 (478) Values Ms1 (478), Ms2 (22)
M_Occup Nominal 0 Least Gover (13) Most other (213) Values other (213), Far_Fish (194), ...[3 more]
M_Income Nominal 0 Least Higher (1) Most Lower (369) Values Lower (369), Non (86), ...[4 more]
P_status Nominal 0 Least Other (4) Most Together (343) Values Together (343), Divorce (103), ...[2 more]
Guardian Nominal 0 Least Relative (63) Most Mothor (299) Values Mothor (299), Father (138), ...[1 more]
Showing attributes 1 - 32
Examples: 500 Special Attributes: 0 Regular Attributes: 32

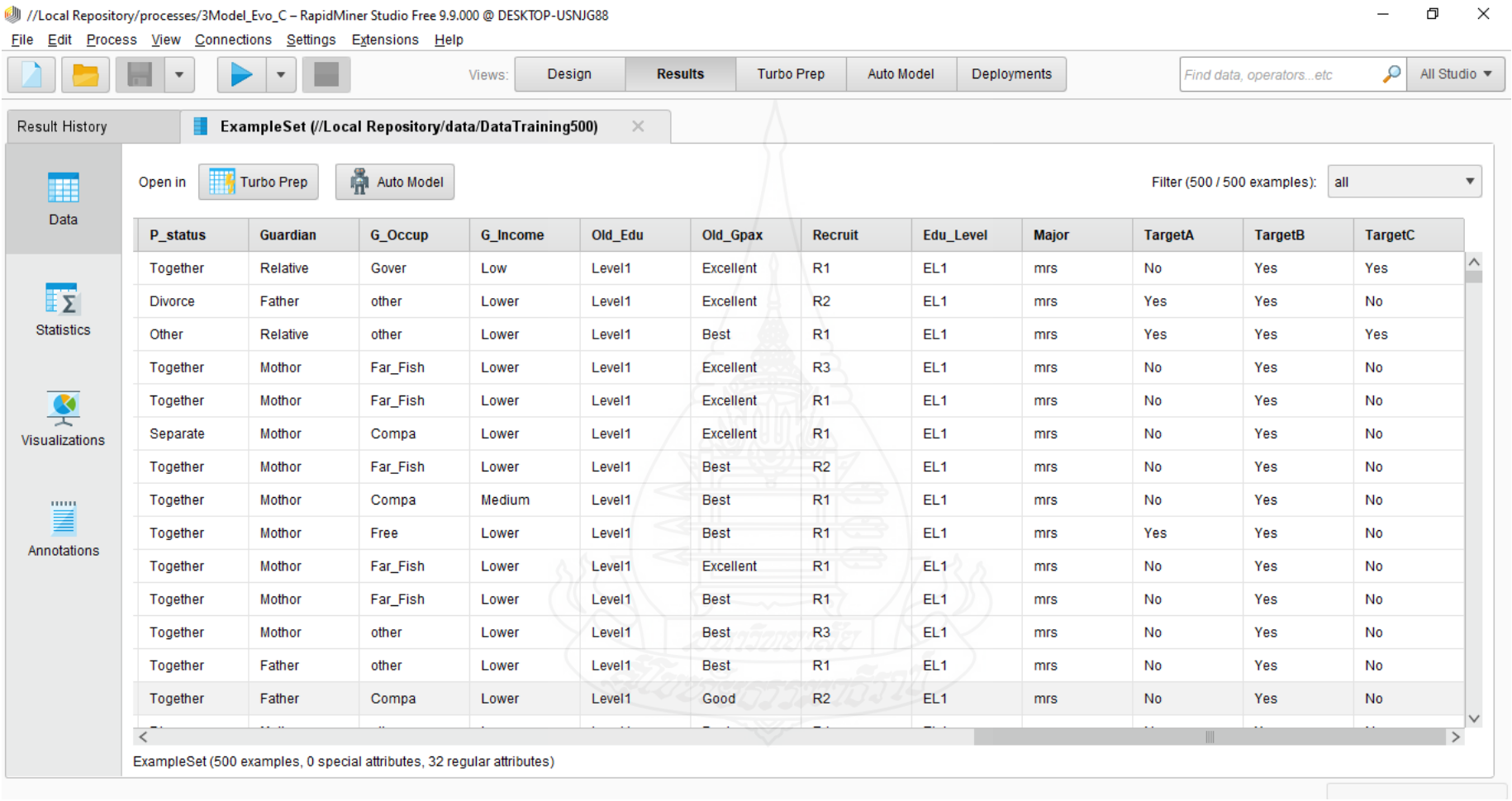

//Local Repository/processes/3Model_Evo_C – RapidMiner Studio Free 9.9.000 @ DESKTOP-USNJG88
File Edit Process View Connections Settings Extensions Help
Views: Design Results Turbo Prep Auto Model Deployments
Find data, operators...etc
All Studio
Result History
ExampleSet (//Local Repository/data/DataTraining500)
Data
Statistics
Visualizations
Annotations
Name Type Missing Statistics
Filter (32 / 32 attributes): Search for Attributes
G_Occup Nominal 0 Least State (2) Most Far_Fish (205) Values Far_Fish (205), other (166), ...[4 more]
G_Income Nominal 0 Least High (2) Most Lower (389) Values Lower (389), Low (60), ...[4 more]
Old_Edu Nominal 0 Least Level1 (500) Most Level1 (500) Values Level1 (500)
Old_Gpax Nominal 0 Least Medium (45) Most Excellent (225) Values Excellent (225), Best (129), ...[2 more]
Recruit Nominal 0 Least R4 (6) Most R1 (308) Values R1 (308), R3 (157), ...[2 more]
Edu_Level Nominal 0 Least EL1 (247) Most EL2 (253) Values EL2 (253), EL1 (247)
Major Nominal 0 Least av (44) Most mrs (315) Values mrs (315), ttm (141), ...[1 more]
Least Most Values
Showing attributes 1 - 32
Examples: 500 Special Attributes: 0 Regular Attributes: 32


//Local Repository/processes/3Model_Evo_C – RapidMiner Studio Free 9.9.000 @ DESKTOP-USNJG88
File Edit Process View Connections Settings Extensions Help
Views: Design Results Turbo Prep Auto Model Deployments
Find data, operators...etc
All Studio
Result History
ExampleSet (//Local Repository/data/DataTraining500)
Data
Statistics
Visualizations
Annotations
Name Type Missing Statistics
Filter (32 / 32 attributes): Search for Attributes
Old_Gpax Nominal 0 Least Medium (45) Most Excellent (225) Values Excellent (225), Best (129), ...[2 more]
Recruit Nominal 0 Least R4 (6) Most R1 (308) Values R1 (308), R3 (157), ...[2 more]
Edu_Level Nominal 0 Least EL1 (247) Most EL2 (253) Values EL2 (253), EL1 (247)
Major Nominal 0 Least av (44) Most mrs (315) Values mrs (315), ttm (141), ...[1 more]
TargetA Nominal 0 Least Yes (102) Most No (398) Values No (398), Yes (102)
TargetB Nominal 0 Least No (40) Most Yes (460) Values Yes (460), No (40)
TargetC Nominal 0 Least Yes (94) Most No (406) Values No (406), Yes (94)
Showing attributes 1 - 32
Examples: 500 Special Attributes: 0 Regular Attributes: 32



Row No.	Gender	Age	Nationality	Race	Religion	Blood	Disease	Disability	Talent	Quota	Hometown
1	F	A1	Thai	Thai	Buddhism	B	No	No	Yes	Healt	Northeast
2	F	A1	Thai	Thai	Buddhism	B	No	No	Yes	Healt	Northeast
3	F	A1	Thai	Thai	Buddhism	B	No	No	No	Healt	Central
4	F	A1	Thai	Thai	Buddhism	O	No	No	Yes	Healt	North
5	F	A2	Thai	Thai	Buddhism	O	No	No	Yes	Healt	Central
6	F	A1	Thai	Thai	Buddhism	B	No	No	Yes	Healt	Northeast
7	F	A1	Thai	Thai	Buddhism	A	No	No	Yes	Healt	Central
8	F	A1	Thai	Thai	Buddhism	A	No	No	No	Healt	Central
9	F	A1	Thai	Thai	Buddhism	B	No	No	No	Healt	Northeast
10	F	A2	Thai	Thai	Buddhism	O	No	No	Yes	Healt	Northeast
11	F	A2	Thai	Thai	Buddhism	B	Yes	No	Yes	Healt	Northeast
12	F	A2	Thai	Thai	Buddhism	O	No	No	Yes	Healt	South
13	F	A1	Thai	Thai	Buddhism	O	No	No	Yes	Healt	South
14	F	A1	Thai	Thai	Buddhism	AB	No	No	Yes	Healt	Northeast


//Local Repository/processes/3Model_Evo_C – RapidMiner Studio Free 9.9.000 @ DESKTOP-USNJG88
File Edit Process View Connections Settings Extensions Help
Views: Design Results Turbo Prep Auto Model Deployments
Find data, operators...etc
All Studio
Result History
ExampleSet (//Local Repository/data/DataTraining500)
Data
Statistics
Visualizations
Annotations
Open in Turbo Prep
Auto Model
Filter (500 / 500 examples): all
address Famsize Son_number F_Status F_Occup F_Income M_status M_Occup M_Income P_status Guardian G_Occup
Country 3 2 Fs1 Gover Medium Ms1 Free Lower Together Relative Gover
City 2 1 Fs1 other Lower Ms1 other Lower Divorce Father other
Country 2 2 Fs2 other Non Ms2 other Non Other Relative other
Country 2 2 Fs1 Far_Fish Lower Ms1 Far_Fish Lower Together Mothor Far_Fish
Country 2 1 Fs1 Far_Fish Lower Ms1 Far_Fish Lower Together Mothor Far_Fish
Country 2 2 Fs1 other Lower Ms1 Compa Lower Separate Mothor Compa
Country 4 4 Fs1 Far_Fish Lower Ms1 Far_Fish Lower Together Mothor Far_Fish
Country 2 2 Fs1 Far_Fish Lower Ms1 Compa Medium Together Mothor Compa
Country 2 2 Fs1 Free Lower Ms1 Free Lower Together Mothor Free
Country 2 2 Fs1 Far_Fish Lower Ms1 Far_Fish Lower Together Mothor Far_Fish
Country 2 1 Fs1 Far_Fish Lower Ms1 Far_Fish Lower Together Mothor Far_Fish
Country 2 2 Fs1 Gover Low Ms1 other Lower Together Mothor other
Country 2 2 Fs1 other Lower Ms1 other Lower Together Father other
Country 1 1 Fs1 Compa Lower Ms1 Far_Fish Lower Together Father Compa
ExampleSet (500 examples, 0 special attributes, 32 regular attributes)



P_status	Guardian	G_Occup	G_Income	Old_Edu	Old_Gpax	Recruit	Edu_Level	Major	TargetA	TargetB	TargetC
Together	Relative	Gover	Low	Level1	Excellent	R1	EL1	mrs	No	Yes	Yes
Divorce	Father	other	Lower	Level1	Excellent	R2	EL1	mrs	Yes	Yes	No
Other	Relative	other	Lower	Level1	Best	R1	EL1	mrs	Yes	Yes	Yes
Together	Mothor	Far_Fish	Lower	Level1	Excellent	R3	EL1	mrs	No	Yes	No
Together	Mothor	Far_Fish	Lower	Level1	Excellent	R1	EL1	mrs	No	Yes	No
Separate	Mothor	Compa	Lower	Level1	Excellent	R1	EL1	mrs	No	Yes	No
Together	Mothor	Far_Fish	Lower	Level1	Best	R2	EL1	mrs	No	Yes	No
Together	Mothor	Compa	Medium	Level1	Best	R1	EL1	mrs	No	Yes	No
Together	Mothor	Free	Lower	Level1	Best	R1	EL1	mrs	Yes	Yes	No
Together	Mothor	Far_Fish	Lower	Level1	Excellent	R1	EL1	mrs	No	Yes	No
Together	Mothor	Far_Fish	Lower	Level1	Best	R1	EL1	mrs	No	Yes	No
Together	Mothor	other	Lower	Level1	Best	R3	EL1	mrs	No	Yes	No
Together	Father	other	Lower	Level1	Best	R1	EL1	mrs	No	Yes	No
Together	Father	Compa	Lower	Level1	Good	R2	EL1	mrs	No	Yes	No

ประวัติผู้วิจัย

ชื่อ	นางสาวนิชาภา จำปาศรี
วัน เดือน ปีเกิด	29 ตุลาคม 2534
สถานที่เกิด	อำเภอโพนพิสัย จังหวัดหนองคาย
ประวัติการศึกษา	วิทยาศาสตรบัณฑิต สาขาเทคโนโลยีสารสนเทศ มหาวิทยาลัยราชภัฏวไลยอลงกรณ์ ในพระบรมราชูปถัมภ์ พ.ศ. 2558
สถานที่ทำงาน	วิทยาลัยเทคโนโลยีทางการแพทย์และสาธารณสุข กาญจนาภิเษก
ตำแหน่ง	ปฏิบัติงานด้านการเรียนการสอน (เทคโนโลยีสารสนเทศ)